UNIVERSITÁ DEGLI STUDI DI MILANO BICOCCA
Corso di Laurea Magistrale in Informatica
FACOLTÁ DI SCIENZE MATEMATICHE FISICHE E NATURALI

An Observational Study: The Effect of Diuretics Administration on Outcomes of Mortality and Mean Duration of I.C.U. Stay

Evolutionary Design and Optimization Group
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

Supervisor: Prof. Giancarlo Mauri
Supervisor: Prof. Leonardo Vanneschi
Supervisor: Prof. Una-May O'Reilly

Master Degree Thesis by:
Daniele Ramazzotti, 725339

Academic Session 2011-2012

To My Parents,
My Grandparents
and My Friends

Acknowledgments

The prospect of carrying out my master thesis at the Massachusetts Institute of Technology (MIT) was proposed me for the first time a year ago by Professor Vanneschi. That possibility was amazing and I immediately decided to accept. My experience at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) began on October 9, 2011 and I stayed in Boston until February 13, 2012 to then return for another five weeks between May and June.

During the 5 months I spent in Boston, I learned an incredible amount of things that allowed me to grow a lot both in professional and personal terms. I can say now that it was an amazing experience that has exceeded even the big expectations I had before leaving.

First of all I would like to thank Professor Leonardo Vanneschi and Professor Una-May O'Reilly for giving me this great opportunity.

I first met Professor Vanneschi during the Soft Computing course. It was really a great experience to learn from him and use the Soft Computing techniques during my stay at the MIT. I would like to thank him the most for offering me the chance to work on my master thesis at the MIT.

I first met professor O'Reilly when I arrived at the CSAIL and during my stay there she taught me a brand new way of working, different from how I was used to. Working with her was incredible.

I would like to thank all the CSAIL team and in particular Dr. James McDermott for helping me to set up and perform the work during the first 4 months of my stay at the CSAIL and Dr. Kalyan Veeramachaneni for helping me in completing the thesis.

I would like to thank Leo Celi M.D. and John Danziger M.D. for their support regarding all the medical issues and the definition of the problem.

I would like to thank my family for always supporting me in my decision to go to Boston and for helping me during all the 5 months.

This experience was incredible. Thank you all for making this possible.

Abstract

This thesis conducts an observational study into whether diuretics should be administered to ICU patients with sepsis when length of stay in the ICU and 30-day post-hospital mortality are considered. The central contribution of the thesis is a stepwise, reusable software-based approach for examining the outcome of treatment vs no-treatment decisions with observational data. The thesis implements, demonstrates and draws findings via three steps:

Step 1. Form a study group and prepare modeling variables.

Step 2. Model the propensity of the study group with respect to the administration of diuretics with a propensity score function and create groups of patients balanced in this propensity.

Step 3. Statistically model each outcome with study variables to decide whether the administration of diuretics has a significant impact.

Additionally, the thesis presents a preliminary machine learning based method using Genetic Programming to predict mortality and length of stay in ICU outcomes for the study group.

The thesis finds, for its study group, in three of five propensity balanced quintiles, a statistically significant longer length of stay when diuretics are administered. For a less sick subset of patients (SAPS ICU admission score < 17) the administration of diuretics has a significant negative effect on mortality.

List of Figures

List of Tables

List of Algorithms

Contents

Chapter 1

Problem Statement

1.1 Introduction

Diuretics are drugs which promote urination. They are often used to bring fluids levels down to normal after intravenous (IV) fluids have been intensively infused. They could be harmful in some circumstances but there are no randomized clinical trials to date which provide evidence for the benefit or harm of these drugs in general. This thesis conducts an observational study into this question.

Generally, all ICU[1] patients with sepsis[2] are infused with high levels of fluids as soon as they enter the ICU in order to improve their low blood pressure and treat their medical condition. This practice is called fluids resuscitation. When a patient is recovering and still has high fluid levels, clinicians face a decision on whether to prescribe diuretics, which will bring about a reduction of a patient's fluids levels to normal, or to let fluids levels decrease naturally. This is a grey area of clinical medicine: different doctors, even when presented with similar patients, can choose either to prescribe diuretics or not.

The long term and broad aim of the developed analysis is to provide clinicians with quantitatively reasoned decision support when they face the choice of treating or not treating. The central contributions include clearly delineated steps describing the analysis see Section 1.3, and a set of software tools. The software tools are re-usable. They support the creation

[1]An Intensive Care Unit (ICU) is a highly specialized department of a hospital that provides intensive-care medicine, concerned with the diagnosis and management of life threatening conditions requiring sophisticated organ support and invasive monitoring.

[2]Sepsis is a potentially deadly medical condition that is characterized by a whole-body inflammatory state and the presence of a known or suspected infection. For a precise description of sepsis see Appendix A.

of a new study group by providing software that extracts, intersects and filters Mimic II Clinical Database records. Additional software supports developing co-variate statistically balanced quintiles of the study group with respect to propensity to receive diuretics. It can be used to develop any propensity score function whether there are treated and untreated patients. Another software module supports outcome modeling with logistic and linear regression accompanied by p-value derivation. The final component of software is slightly modified genetic programming-based machine learning code which executes symbolic regression for classification and regression plus code calling a library function that performs clustering, a form of unsupervised learning. Specifically, in this thesis, the aim is to retrospectively examine the data of ICU patients with sepsis and, while controlling for the propensity for the administration of diuretics and considering the possibly **confounding factor** of illness, to determine whether the administration of diuretics has a significant effect on 30 day mortality outcome post-ICU or mean duration of ICU stay.

1.2 Study Group

The analysis attempts to address a specific group of patients, in particular adult patients with a large amount of fluids in their bodies. Therefore, the analysis will be conducted on patients over 18 years old[3] and with a sepsis diagnosis[4]. From this group, CMO[5] patients have been filtered out because their outcome with respect to mortality is distinctive. Patients who had been taking diuretics before entering the ICU have also been eliminated because of the compliating nature of this on a decision for the administration of diuretics (or continuing to take them). Finally, patients who had multiple admissions, both in ICU and in the hospital have been eliminated.[6].

1.2.1 Variables of the Outcome Study and/or Propensity Model

A certain number of variables have been used as variables in the study to describe the condition of a patient during his/her stay in the ICU. Now a

[3]Neonatal sepsis is not subject of study in this work.

[4]It is difficult to define sepsis. In this work, as described in Appendix A, have been used the definition described in[1].

[5]Comfort Measures Only refers to medical treatment of a dying person where the natural dying process is permitted to occur while assuring maximum comfort.

[6]First ICU visit data is a good potential alternative filter in a follow up study group creation.

brief medical description of those factors is presented. Much of the text is directly quoted because of the need for medical precision.

- SAPS II score: this point score is based upon a severity of disease classification system[2]. It is calculated from 12 routine physiological measurements during the first 24 hours, information about previous health status and some information obtained at hospital admission.

- SOFA score: this point score is based upon a scoring system to determine the extent of a person's organ function or rate of failure[3]. The score is based on six different scores, one each for the respiratory, cardiovascular, hepatic, coagulation, renal and neurological systems.

- Elixhauser score: this score integrates a list of 30 comorbidities relying on the ICD-9-CM coding manual. The comorbidities were not simplified as an index because each comorbidity affected outcomes (length of hospital stay, hospital changes, and mortality) differently among different patients groups. The comorbidities identified by the Elixhauser comorbidity measure are significantly associated with in-hospital mortality and include both acute and chronic conditions. Walraven et al.[4] has derived and validated an Elixhauser comorbidity index that summarizes disease burden and can discriminate for in-hospital mortality.

- Creatinine: this is a break-down product of creatine phosphate in muscle. It is usually produced at a fairly constant rate by the body (depending on muscle mass). In our study, this factor is included because it can be indicative of kidney disease.

- (Administration of) Vasopressors: Vasopressors indicates whether the patient was administered any sort of vaso-suppressor. Vasopressors are drugs that constrict the blood vessels and thereby elevate blood pressure. Usually, before a patients is considered able to leave the ICU, vasopressors are suspended. The administration of any vaso-suppressor have been included as a factors because vaso-suppressors are indicative of health condition.

- Mechanical ventilation: this boolean variable indicates whether or not the patient was mechanically ventilated. Mechanical ventilation assists or replaces spontaneous breathing. Ventilation may involve a machine called a ventilator or the breathing may be assisted by a physician, respiratory therapist or other suitable person compressing a bag or set of bellows.

- Arterial blood pressure: this quantity is the pressure exerted by circulating blood upon the walls of blood vessels and is one of the principal vital signs. The blood pressure in the circulation is principally due to the pumping action of the heart. In the study, the value for arterial blood pressure refers to systolic[7].

- Mean arterial blood pressure: This quantity is defined as arterial pressure during a single cardiac cycle. Is is calculated as $\frac{2 \cdot diastolic + systolic}{3}$.

1.3 Analysis Steps

There are 4 steps in the developed statistical analysis.

Step 1: Form a study group and prepare modeling variables. In this step, subsets of patient records are extracted from the Mimic II Clinical Database and the subsets are fused then filtered according to clinician input. The variables of the study group are next prepared for subsequent analysis. This step is described in Chapter 2.

Step 2: Model the propensity of the study group with respect to the administration of diuretics with a propensity score function and create groups of patients balanced in this propensity. This step is described in Chapter 3.

Step 3: Statistically model outcome with study variables to decide whether the administration of diuretics has a significant impact on mortality and length of stay in ICU while considering health condition and propensity of the administration of diuretics. This step is described in Chapter 4.

Step 4: Design a preliminary machine learning based method using Genetic Programming to predict mortality and length of stay in ICU outcomes for the study group. This step is described in Chapter 5.

An overview of this process is shown in Figure 1.1 on the facing page. Finally, Chapter 6 summarizes the analysis and its findings and lists possible future work.

[7]During each heartbeat, blood pressure varies between a maximum (systolic) and a minimum (diastolic) pressure. Systolic blood pressure is a measure of blood pressure while the heart is beating, while diastolic pressure is a measure of blood pressure while the heart is relaxed.

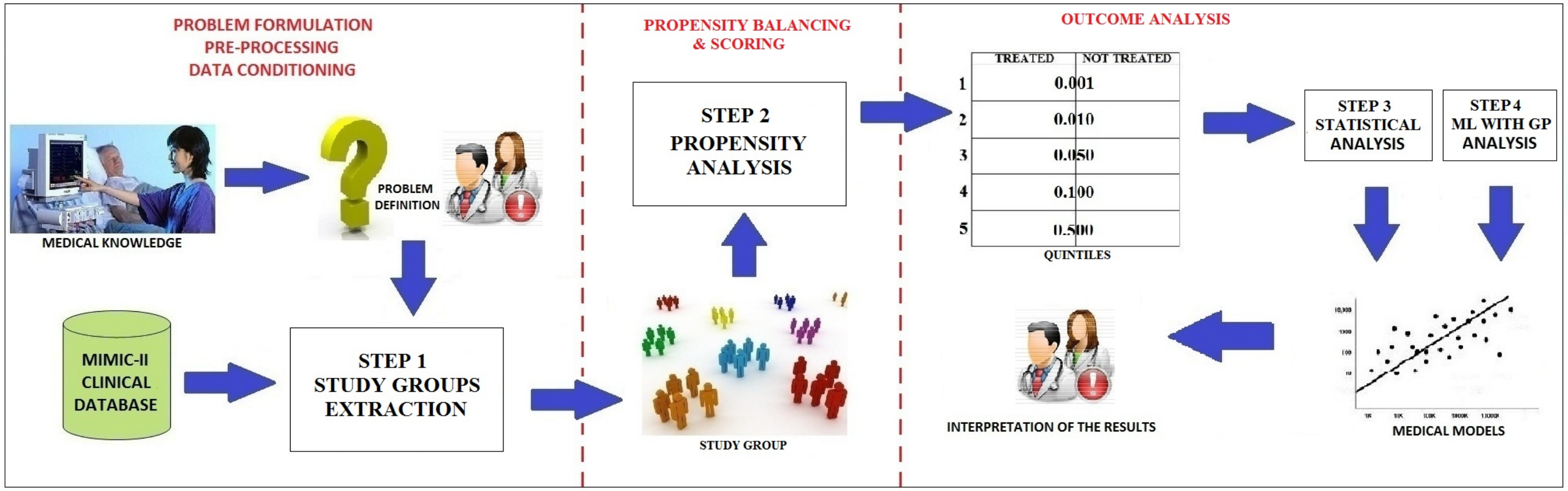


Figure 1.1: The workflow used to define and solve the problem. The definition of the problem and in particular of the Study Group required a lot of work in collaboration with medical experts. The process was subject to a progressive refinement at the end of which the dataset extraction was performed.

Chapter 2

Study Group Extraction

The aim of this Chapter is to show Step 1 of the analysis in which the study group is identified by means of extraction, intersection and filtering of records from the Mimic II Clinical Database and the variables are conditioned for subsequent propensity balancing and outcome analysis modeling.

In Figure 2.1 on the next page the software modules supporting Step 1 are shown. The output after dataset extraction from the Mimic II Clinical Database is a series of flat files. In the dataset preprocessing module, the flat files provided by the dataset extraction were then merged into a flat file containing all the data. A detailed description of the contributed software is provided in Appendix B.

The chapter proceeds in the following manner: Section 2.1 starts with a description of the Mimic II Clinical Database. All the contents of this Section are drawn from [5] and a deeper description of the database can be found there. A description of how the study group was formed via a series of extractions, intersections and filters then follows in Section 2.2. Descriptive statistics on the study group is also provided. Section 2.3 explains how each of the variables for the modeling steps were prepared. Timeline oriented variables were worthy of explicit attention. It provides a complete list of every variable prepared for modeling and a breakdown of how many patients the administration of diuretics vs those not were distributed for the variable.

2.1 Mimic II Clinical Database

The Mimic II Clinical Database (Multiparameter Intelligent Monitoring in Intensive Care) records data from all ICU patients in the Beth Israel Deaconess Medical Center. It is notable for three factors: it is publicly and freely available; it encompasses a diverse and very large population of ICU

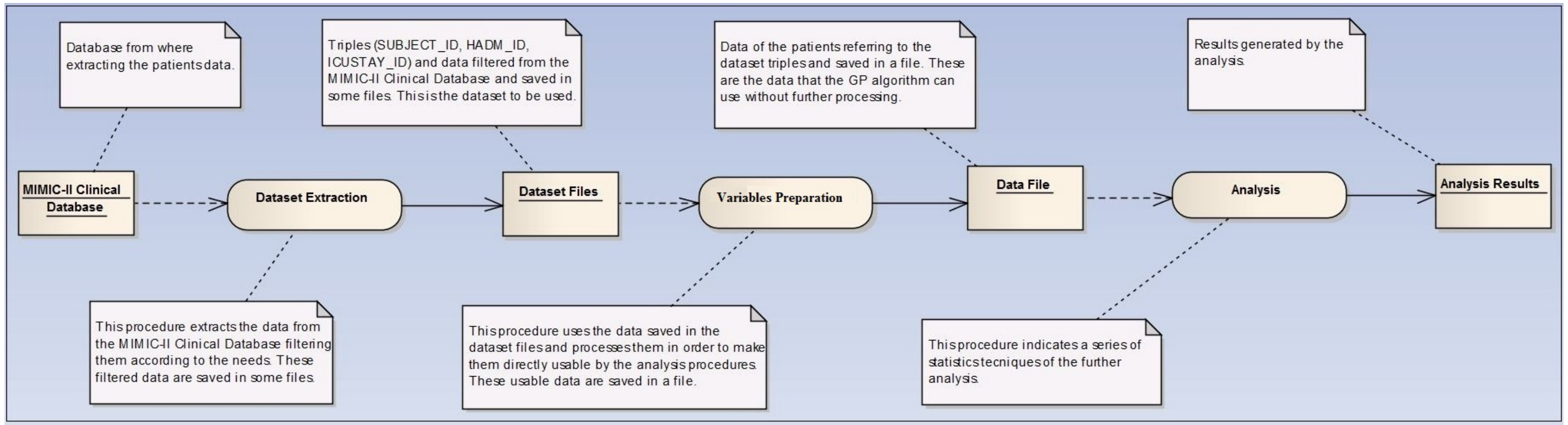


Figure 2.1: The whole process of analysis in details. After the extraction, the variables preparation was performed to save all the needed data.

patients; and it contains high temporal resolution data including lab results, electronic documentation, and bedside monitor trends and waveforms. The database can support a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development.

The process and the sources of the data collection for the Mimic II Clinical Database is shown in Figure 2.2. The data is collected dating from

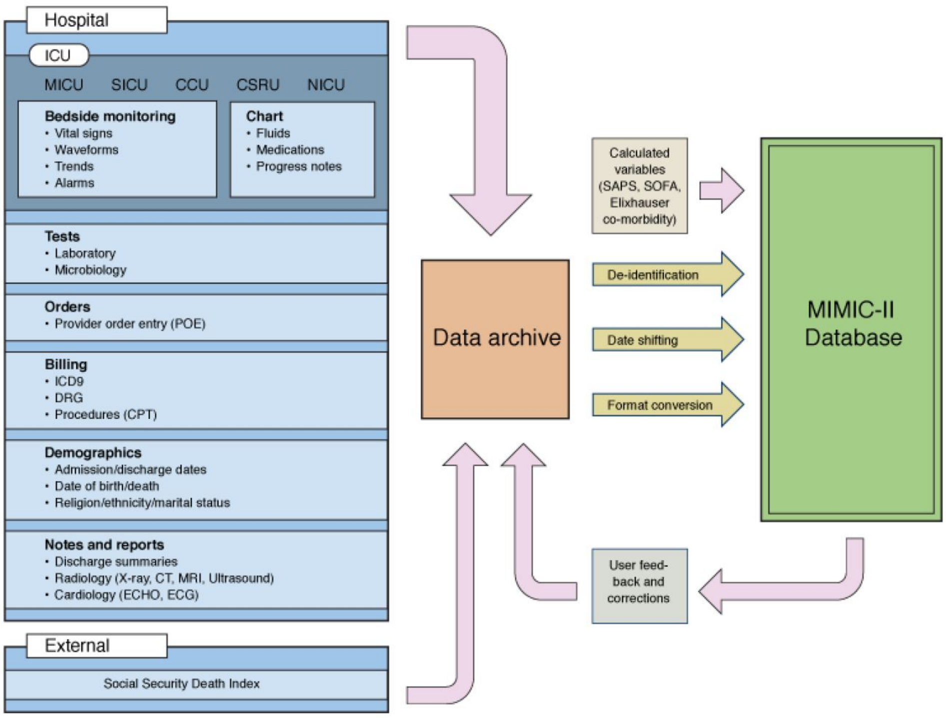


Figure 2.2: Schematic of data collection and database construction. Source data consists of: bedside monitor waveforms and trends, the ICU clinical databases, the hospital archives and the Social Security Death Index. These data are assembled in a protected and encrypted database which is then de-identified to provide one relational database plus associated flat file bedside waveforms and trends.

2001 from Boston's Beth Israel Deaconess Medical Center (BIDMC). Any patient who was admitted to the ICU on more than one occasion may be represented by multiple patient visits. The adult ICUs (for patients aged 15 years and over) include medical (MICU), surgical (SICU), coronary (CCU), and cardiac surgery (CSRU) care units. Data were also collected from the neonatal ICU (NICU).

Clinical data are recorded far less frequently than bedside monitor data and come from a variety of databases. These include the laboratory results, pharmacy provider order entry (POE records, admission and death

records, demographic details, discharge summaries, ICD-9 codes, procedure codes, microbiology and lab tests, imaging and ECG reports and the ICU central database (which includes some subset of the bedside monitor trends, drip rates, free text nursing notes and nurse-verified down-sampled trends, amongst other information).

2.1.1 Definition of Patient Record

Since a patient may have been admitted several times during the period in which our data were collected, it is important to understand exactly how to identify patients and his/her stay(s).

There are essentially four identifiers for data associated with any given patient:

- **SUBJECT_ID:** to identify the patient. It is an integer number identifying a particular patient. This can be thought of as a substitute for a unique medical record number. In the flat file data posted on PhysioNet, the number representing the SUBJECT_ID is left padded with zeros to five digits and preceded by the letter s. In the relational database, the SUBJECT_ID has no preceding letter or leading zeros.
- **HADM_ID:** to identify the admission in the hospital. It is an integer number identifying a particular admission to the hospital. Each patient may have many HADM_IDs associated with his/her unique SUBJECT_ID.
- **ICUSTAY_ID:** to identify the admission in the ICU. It is an integer number identifying an ICU stay. An ICU stay, refers to the period of time when the patient is cared for continuosly in an Intensive Care Unit. Each patient may have one or more ICU stays associated. An ICU stay is considered to be continuous if any set of ICU events (such as bed transfers or changes in type of service) belonging to one SUBJECT_ID which are fewer than 24 hours apart. Longer breaks in the patient's stay automatically cause a new ICUSTAY_ID to be assigned.

Figure 2.3 on page 12 illustrates the possible data available for a given individual, identified by a ICUSTAY_ID. Time progresses from left to right, and the different types of data collected are shown vertically. Each subject can have multiple hospital admissions, identified with HADM_IDs. Each hosptial admission can contain multiple ICU stays, identified with ICUSTAY_IDs. Laboratory and microbiology tests are performed throughout a hospital stay and can therefore take place outside the ICU stay. Vital sign

validation, medications, fluid balances and nursing notes are only performed in the ICU and are not available during the remainder of the hospital stay. Date of death is recorded in-hospital and has also been obtained from social security records for out-of-hospital mortality. The above illustrates an ideal case where the timestamps associated with the data fall within the hospital and/or ICU stay. Unfortunately, real-world issues can complicate matters allowing data to be recorded outside of a patient stay. For example, a patient could be physically present in the ICU and connected to monitors before their admission has been entered into the system. This results in a waveform recording which starts before the subject's ICU admission. Furthermore, missing/mistaken data can mean that ICU stays exist where there is no matching hospital admission record.

Note that a patient may move between ICUs during any given admission. If the move is longer than 24 hours, it is defined to be a new ICU stay. Note also that the amount of data varies during and between ICU stays and that data are often missing.

The Mimic II Clinical Database is a relational database.

2.2 Dataset Extraction

The study group was extracted with a series of extractions, intersections and filters.

2.2.1 Extractions, Intersections and Filtering

Step 1 consists of a series of queries to the Mimic II Clinical Database. These queries extract the data to 21 files and generate record sets. Of the 21 extracted files, the first 3 contain the triples (SUBJECT_ID, HADM_ID, ICUSTAY_ID). These triples, which identify respectively the patients, the admission to the hospital and the admission in the ICU, identify uniquely the data for each feature extracted in the remaining files.

An other file is generated which contains the patient *discharge summaries*. They are reports in a English text-like format where the clinical history of the patients before and during the stay in the hospital is described.

Finally, the last 17 extracted files contain the variables of interest for each patient.

In the Mimic II Clinical Database there are 39,919 ICUSTAY_ID records. The impact on the filter is shown for each step. The values referred as *original db* report how the single step impacts on the Mimic II Clinical

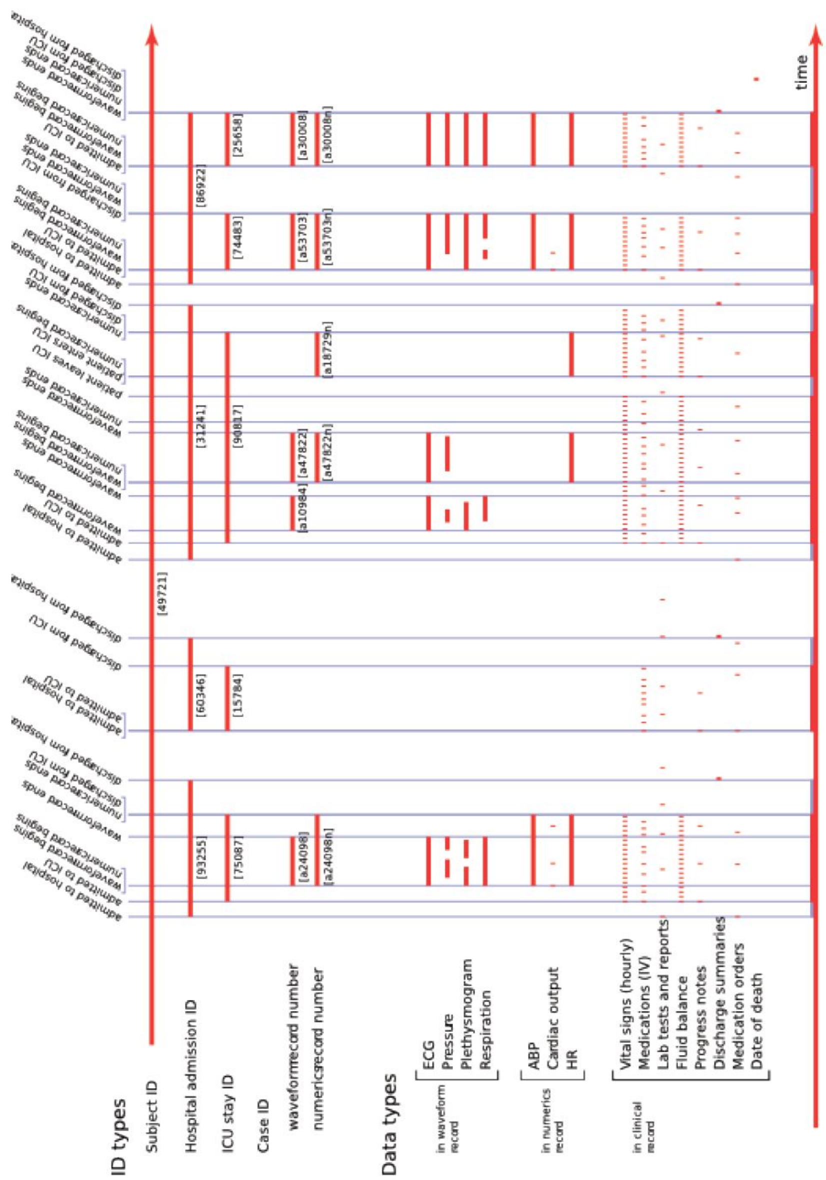


Figure 2.3: Schematic of a patient record. Note that the patient may experience several hospital admissions and ICU stays, for which differing amounts of data are available.

Database. The values referred as *previous subset* report how the single step impacts on the subset after the previous step. The values referred as *all conditions* report how the interactions between all the steps applied till that point impact on previous step. The sequence of filtering steps follow:

1. Extract patients with all the 3 IDs (SUBJECT_ID, HADM_ID, ICUSTAY_I(d) available and different from null:
   - 36,708/39,919 91.95% (original db)
2. Extract patients with only one admission in the hospital and in the ICU:
   - 26,027/39,919 65.19% (original db)
3. Intersect 1 and 2:
   - 26,027/36,708 70.90% (previous subset)
   - 26,027/39,919 65.19% (all conditions)
4. Extract patients with at least 1 full day of data in the ICU:
   - 29,462/39,919 73.80% (original db)
5. Intersect 3 and 4:
   - 18,770/26,027 72.11% (previous subset)
   - 18,770/39,919 47.02% (all conditions)
6. Extract patients with age equal or over 18. The aim of doing this is to remove neonates:
   - 31,859/39,919 79.80% (original db)
7. Intersect 5 and 6:
   - 15,176/18,770 80.85% (previous subset)
   - 15,176/39,919 38.01% (all conditions)
8. Extract patients with sepsis according to [1]:
   - 3,818/39,919 9.56% (original db)
9. Intersect 7 and 8:
   - 1,644/15,176 10.83% (previous subset)

   - 1,644/39,919 4.11% (all conditions)

10. Extract patients not CMO (comfort measure only):
    - 39,651/39,919 99.32% (original db)
11. Intersect 9 and 10:
    - 1,644/1,644 100% (previous subset)
    - 1,644/39,919 4.11% (all conditions)
12. Extract patients with a discharge summary available:
    - 31,475/39,919 78.84% (original db)
13. Intersect 11 and 12:
    - 1,638/1,644 99.63% (previous subset)
    - 1,638/39,919 4.10% (all conditions)
14. Extract patients who are diuretics naive. See note in 2.2.1.1:
    - 30,646/39,919 76.77% (original db)
15. Intersect 13 and 14:
    - 1,606/1,638 98.04% (previous subset)
    - 1,606/39,919 4.02% (all conditions)
16. Filter 15 for missing data:
    - 1,606/1,638 94.76% (previous subset)
    - 1,522/39,919 3.81% (all conditions)

For a detailed description of the results of this step see Appendix C.

2.2.1.1 Diuretics Naive Status

The study was intended to consider patients who had not been administered diuretics before entering the ICU. This is intended to avoid having data which would be conditioned before the interval of study. Unfortunately this information is not directly available from the Mimic II Clinical Database. It needs to be parsed out of the discharge summary. The discharge summary is saved in the database in an English text-like format. It consists of a summary of what has happened to the patient before and during his admittance at

the hospital. The document is hand written by clinicians so it does not have a well defined structure. The needed pieces of information were extracted using a complicated parser which searched the English text for the names of a list of diuretics decided by the doctors. Those patients who had not been administered diuretics before entering the ICU were considered *diuretics naive*.

2.2.2 Filters

Step 2 takes as an input the files provided by the SQL queries and the list of diuretics naive patients. The aim of this step is to combine the results of the SQL queries to the one of the diuretics naive procedure. Moreover, the procedure discards the patients who have missing data for any variable needed for the analysis.

The output of this phase is a series of files where all the available data are saved. A file per variable is then created.

- Results 1,522/39,919 patients (3.81%)
- To 189/1,522 patients (12.41%) diuretics have been given, D^+
- To 1,333/1,522 patients (87.59%) diuretics have not been given, D^-

A summary of all the steps of the dataset extraction are shown in Table 2.1 on the following page. The final number of patients considered for the study is 1,522.

2.3 Variables Preparation

Critical care medical data is arguably the most valuable clinical data supporting medical informatics. This is because the ICU is the crucible of a hospital. It accepts the most acutely ill of patients, it uses pervasive monitoring, and intensivists encounter frequent medical episodes for which they must make rapid interventions.

ICU medical doctors, specialists known as intensivists, are key members of any knowledge mining team which consults a data resource such as the Mimic II Clinical Database. Data engineers, with expertise in modeling and machine learning, request a lot of information from a team's intensivists. A key request they make is the variables that should be selected for predictive or explanatory models which will be mined from the data.

For example, in a study on the necessity of so called *diuretics* (drugs) for diuresis (fluid shedding), after fluid resuscitation in the ICU, it must

Step	Type	Effects	%
1	Extract(A)	36,708	91.95%
2	Extract(B)	26,027	65.19%
3	$A \cap B \rightarrow C$	26,027	65.19%
4	Extract(D)	29,462	73.80%
5	$C \cap D \rightarrow E$	18,770	47.02%
6	Extract(F)	31,859	79.80%
7	$E \cap F \rightarrow G$	15,176	38.01%
8	Extract(H)	3,818	9.56%
9	$G \cap H \rightarrow I$	1,644	4,11%
10	Extract(L)	39,651	99.32%
11	$I \cap L \rightarrow M$	1,644	4.11%
12	Extract(N)	31,475	78.84%
13	$M \cap N \rightarrow O$	1,638	4.10%
14	Extract(P)	30,646	76.77%
15	$O \cap P \rightarrow Q$	1,606	4.02%
16	Filter(Q) $\rightarrow$ R	1,522	3.81%

Table 2.1: Summary of the 16 steps of the dataset extraction. Next of each step, is shown how it impacts on the Mimic II Clinical Database. The number of patients in the final subset is 1,522.

be determined what intravenous diuretics data should be included as model variables.

Indicator selection calls upon intensivists' theoretical and clinical knowledge and their ICU experience. The selection process is intensely deliberative and uncertain. The intensivists recognize how the events of each patient's ICU stay are unique and how the care administered is both subjective and informed by medical knowledge. They are uncertain as to whether some of the collected data variables confound the outcome that is to be explained. As well, they are aware that among intensivists there exists a propensity to treat similar patients differently. They find it very challenging to choose variables that essentially pinpoint some time point or variable of a complex human health process. For example, while including an feature expressing the amount of a diuretic is an obvious decision, how the amount is described is open to debate.

The uncertainty of the decisions imply that the feature selection process is challenged to be systematic and unbiased and impacts the quality and accuracy of modeling in an obviously critical way. In addition to the uncertainty, from a broader perspective, the selection of variables prior to their experimentation in model regression is problematic. It forces a model design decision that is premature because the appropriate information is unavailable.

In this work, the definition of the problem and descriptive variables to be included in the model was carried out through interviews and meetings with a group of doctors, experts in the field. In particular, from an initial definition of the problem due to the intuition and experience of the medical experts, through this process a more formal definition has been gained of all variables needed to define a comprehensive model. This process went hand in hand with an increasing acquisition of medical knowledge needed to define, from a statistical point of view, an effective model that describes the problem. Another fundamental endeavor has been made in understanding the Mimic II Clinical Database and the related problems that occurred in the preprocessing step.

2.3.1 Timelines

Some of the variables have timelines, this means that for those variables there are values at different times. It is difficult to choose good times where to take the values avoiding forward-looking variables. In the study the clinical data of the patients are available during the whole ICU stay, but the doctors, while making their decision whether giving or not diuretics,

are considering only the values till the current day. In this sense, values available after the diuretics decision points are considered forward-looking variables, in fact their values may be caused by the diuretics decision itself.

To avoid such problems, for those variables the extraction have been performed at the following times:

- **T1** *Diuretics Decision Timepoint*: This is when the decision to give diuretics was 'theoretically' or actually taken. There are two possibilities:
  - for a patient who got diuretics, the actual time of the first dose;
  - for the patients who didn't get diuretics, the **T1** timepoint is day 4. The decision to use day 4 was reached by examining the patients who got diuretics in the first week of their ICU stay. Among this group the first day they were administered diuretics was examined. From this, 'first administration day', data, the median was extracted (day 4) as the timepoint.
- **T2** *Max Fluids Ratio Timepoint: Day of highest fluids ratio*: Ideally this timepoint would be determined individually for each member of the study group. It would express variables at the time when a patient has his/her highest fluids ratio. However, this would have required us to look forward in the data which would make calculation of the timepoint in reality impossible. Instead, day 3 in the ICU T2 has been selected. It was selected by examining the day of highest fluids ratio for all patients in the study group and choosing the median. The fluids ratio is calculated as

$$\frac{inputs(t-1) + inputs(t)}{outputs(t-1) + outputs(t)} \tag{2.1}$$

- **T3** *2nd Highest Fluids Ratio Timepoint: Day of 2nd highest fluids ratio*: The timepoint of second highest fluids ratio has been added because the day of highest fluids ratio is typically very close to the first day of ICU admission when fluids are infused. It may reflect recent infusion more than a delay in shedding fluid.

  Ideally this timepoint would be determined individually for each member of the study group. It would express variables at the time when a patient has his/her second highest ratio of fluids in the body. However, this would have required us to look forward in the data which would make calculation of the timepoint in reality impossible. Instead, T3

have been selected to be day 4. This day was selected by examining the day of the second highest fluids ratio for all patients in the study group and choosing the median.

Figure 2.4 provides a visual summary of the considered times and of the ICU timeline in general. In Appendix C are available more details on how these times were chosen.

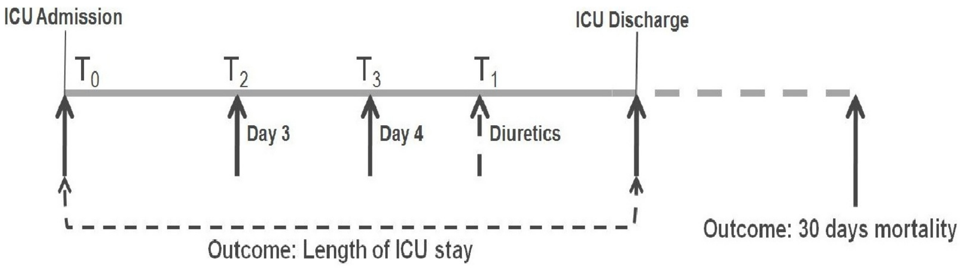


Figure 2.4: The timepoints where the values for timeline variables were acquired. All the values refer to the ICU stay. T0 is day 1, T2 and T3 are days 3 and 4. T1, the diuretics decision time, is day 4 for the patients who didn't get diuretics and it is the actual day when the drug was first administered in the patients who got diuretics. The choice of these timespoint allows us to avoid forward-looking variables.

2.3.2 List of Variables

The flat files produced by the previous modules are the input of a series of procedures which elaborate and save them in a format directly usable as an input for the further analysis. As has been said, the inputs of these modules are a series of files provided by the previous phase, then on those files are computed the time points as discussed in 2.3.1.

In the Mimic II Clinical Database the values are saved in an irregular sampling rate, each values is recorded when available tipically a few times each days. As for this study the values were needed on a daily basis, the original irregular sample rate has been downsampled to fit a regular daily rate of values. This has been done by computating the median value of the available values each day. The median value also made the system robust to outlayers.

Then the list of variables are generated. In the following list is reported the type of the feature, and where needed a brief description:

1. Diuretics in the ICU:

   x_1 Binary: -1 not given, +1 given;

2. Age when admitted in the ICU:

   x_2 Numeric;

3. Gender:

   x_3 Binary: -1 male, +1 female;

4. Race (white vs not white):

   x_4 Binary: -1 not white, +1 white;

5. Saps:

   x_5 Average from day 1 to day T1;

   x_6 Mean of values during the first day;

   x_7 Mean of values during day T1;

   x_8 Mean of values during day T2;

   x_9 Mean of values during day T3;

6. Sofa:

   x_{10} Average from day 1 to day T1;

   x_{11} Mean of values during the first day;

   x_{12} Mean of values during day T1;

   x_{13} Mean of values during day T2;

   x_{14} Mean of values during day T3;

7. Elixhauser overall[1].

   x_{15} Numeric;

8. Elixhauser binary: 9 of the 30 fields composing the Elixhauser score are selected: congestive heart failure, cardiac arrhythmias, valvular disease, hypertension, diabetes uncomplicated, diabetes complicated, renal failure, liver disease and obesity.

   $x_{16} \rightarrow x_{24}$ Binary: -1 not present, +1 present;

9. Creatinine:

   x_{25} Average from day 1 to day T1;

[1] The sum of all the parameters of the Elixhauser score, as explained in Chapter 1.2.1.

x_{26} Mean of values during the first;

x_{27} Mean of values during day T1;

x_{28} Mean of values during day T2;

x_{29} Mean of values during day T3;

10. Fluids inputs in liters:

    x_{30} Average of sums from day 1 to day T1;

    x_{31} Sum of values during the first day;

    x_{32} Sum of values during day T1;

    x_{33} Sum of values during day T2;

    x_{34} Sum of values during day T3;

11. Fluids outputs in liters:

    x_{35} Average of sums from day 1 to day T1;

    x_{36} Sum of values during the first day;

    x_{37} Sum of values during day T1;

    x_{38} Sum of values during day T2;

    x_{39} Sum of values during day T3;

12. Fluids balance in liters (fluids inputs - fluids outputs):

    x_{40} Average of sums from day 1 to day T1;

    x_{41} Sum of values during the first day;

    x_{42} Sum of values during day T1;

    x_{43} Sum of values during day T2;

    x_{44} Sum of values during day T3;

13. Use of vasopressors in the ICU:

    x_{45} Binary: -1 not given, +1 given;

14. Mechanical ventilation in the ICU:

    x_{46} Binary: -1 not happened, +1 happened;

15. Arterial blood pressure:

    x_{47} Average from day 1 to day T1;

    x_{48} Mean of values during the first day;

x_{49} Mean of values during day T1;

x_{50} Mean of values during day T2;

x_{51} Mean of values during day T3;

16. Mean arterial blood pressure:

x_{52} Average from day 1 to day T1;

x_{53} Mean of values during the first day;

x_{54} Mean of values during day T1;

x_{55} Mean of values during day T2;

x_{56} Mean of values during day T3;

17. Mortality within 30 days:

x_{57} Binary: -1 alive, +1 dead;

18. Length of stay in the ICU after the first dose of diuretics:

x_{58} Numeric, in days;

For the medical meaning of some of those variables, see Chapter 1.2.1. Table 2.2 on the facing page shows a summary for the values for the binary variables.

Tables 2.3 on the next page, 2.4 on page 24 and 2.5 on page 25 show a summary of the values for the not binary variables, for the variables with timeline which refer to the clinical condition of the patient and for the variables with timeline which refer to the fluids measurements.

Detailed descriptions of those variables and of sepsis are available respectively in Chapter 1.2.1 and Appendix A. Instead for a more detailed description of the procedures produced to realize the study group extraction see Appendix B.

Variable		D^+ (189)		D^- (1333)	
Name	i	# Positive	%	# Positive	%
DIU	1	189	100%	-	-
GEN	3	81	42%	566	42%
RAC	4	1	0.5%	29	2%
EL1 (CHF)	16	91	48%	418	31%
EL2 (Cardia arrythmia)	17	70	37%	335	25%
EL3	18	26	13%	110	8%
EL4	19	51	26%	366	27%
EL5	20	47	24%	258	19%
EL6	21	9	4%	77	5%
EL7	22	10	5%	115	8%
EL8	23	16	8%	125	9%
EL9	24	5	2%	18	1%
VAS (vasosuppressor)	45	163	86%	865	64%
VEN (ventilation)	46	177	93%	904	67%
MOR	57	63	32%	504	37%

Table 2.2: Descriptive statistics for unbalanced, binary variables in the study group. The abbreviation of the variable, its x_i subscript, the number of patients positive for the variable within the patients who got diuretics and within the patients who did not get diuretics are reported. The 4 variables where there is a noticeable, major difference between the diuretics treated and non-treated patients have been highlighted however the study group is not balanced in terms of covariates.

Variable		$D^+ \cup D^-$ (1522)		D^+ (189)		D^- (1333)	
Name	i	Mean(SD)	Me	Mean(SD)	Me	Mean(SD)	Me
AGE	2	66.1(16.9)	68.3	66.2(15.4)	68.9	66.1(17.1)	68.2
ELI	15	3(1.6)	3	3.1(1.4)	3	2.9(1.6)	3
LOS	58	7.4(11.8)	3	15.1(14)	11	6.3(11)	3

Table 2.3: Descriptive statistics for unbalanced discrete variables in study group. The abbreviation of the variable, its x_i subscript, the mean, standard deviation and median of the variable within the study group, patients who got diuretics and within the patients who did not get diuretics are reported. There is a noticeable difference between the length of stay of diuretics treated and non-treated patients which has been highlighted however the study group is not balanced in terms of covariates.

Variable		$D^+ \cup D^-$ (1522)		D^+ (189)		D^- (1333)	
Name	i	Mean(SD)	Me	Mean(SD)	Me	Mean(SD)	Me
SA1	5	15.8(4.6)	15.4	16.2(3.4)	16	15.7(4.8)	15.3
SA2	6	17.1(5.4)	17	17.9(4.9)	18	17(5.5)	17
SA3	7	15(5.2)	15	15.7(4.3)	16	14.9(5.3)	15
SA4	8	15.3(5.2)	15	16.3(4.6)	16	15.2(5.2)	15
SA5	9	17.1(5.4)	17	17.9(4.9)	18	17(5.5)	17
SO1	10	8.3(4.2)	7.6	9.7(3.6)	9.6	8.1(4.3)	7.4
SO2	11	9.1(4.5)	9	10.3(4.2)	11	8.9(4.5)	8
SO3	12	7.6(4.8)	7	9.5(4.3)	9	7.3(4.9)	7
SO4	13	8.2(4.7)	7.4	10(4.1)	10	7.9(4.7)	7
SO5	14	9.1(4.5)	9	10.3(4.2)	11	8.9(4.5)	8
CR1	25	1.8(1.8)	1.2	1.6(1.5)	1.2	1.8(1.8)	1.2
CR2	26	1.8(2.2)	1.2	1.5(1.3)	1.1	1.8(2.3)	1.3
CR3	27	1.8(2)	1.2	1.6(1.4)	1.1	1.8(2.1)	1.2
CR4	28	1.7(1.5)	1.2	1.6(1.4)	1.2	1.8(1.6)	1.2
CR5	29	1.8(2.2)	1.2	1.5(1.3)	1.1	1.8(2.3)	1.3
BP1	47	113(17.3)	114	113(14.7)	113	113(17.7)	114
BP2	48	110(18.7)	114	108(19.5)	106	110(18.5)	114
BP3	49	114(22.1)	114	114(19.2)	112	114(22.5)	114
BP4	50	113(19.6)	114	111(17.7)	110	113(19.8)	114
BP5	51	110(18.7)	114	108(19.5)	106	110(18.5)	114
BM1	52	78.6(12.5)	79.2	77.2(8.8)	77	78.8(13)	79.2
BM2	53	77.6(14.6)	79.2	74.8(12.1)	74	78(14.9)	79.2
BM3	54	78.9(15.3)	79.2	77.1(12.8)	75	79.2(15.6)	79.2
BM4	55	78.4(13.9)	79.2	75.5(11.2)	73	78.8(14.2)	79.2
BM5	56	77.6(14.6)	79.2	74.8(12.1)	74	78(14.9)	79.2

Table 2.4: Descriptive statistics for unbalanced variables on timepoints (part 1). The abbreviation of the variable, its x_i subscript, the mean, standard deviation and median of the variable within the study group, patients who got diuretics and within the patients who did not get diuretics are reported.

Variable		$D^+ \cup D^-$ (1522)		D^+ (189)		D^- (1333)	
Name	i	Mean(SD)	Me	Mean(SD)	Me	Mean(SD)	Me
FI1	30	1.7(1.3)	1.4	1.4(1.2)	1.1	1.7(1.3)	1.4
FI2	31	2.6(2.3)	2	3(3.2)	2.2	2.5(2.1)	2
FI3	32	1.1(1.2)	0.7	1(1.1)	0.6	1.1(1.2)	0.7
FI4	33	1.4(1.4)	1	1.6(1.7)	1	1.4(1.4)	1
FI5	34	2.6(2.3)	2	3(3.2)	2.2	2.5(2.1)	2
FO1	35	1.5(1.1)	1.3	1.8(1.5)	1.6	1.4(1.1)	1.2
FO2	36	1.6(1)	1.2	1.6(3.4)	1.1	1.5(1.4)	1.2
FO3	37	1.4(1.3)	1.1	1.9(1.7)	1.7	1.3(1.2)	1
FO4	38	1.4(1.5)	1.1	1.6(3.1)	1.1	1.3(1.1)	1.2
FO5	39	1.6(1.8)	1.2	1.6(3.4)	1.1	1.5(1.4)	1.2
FB1	40	0.2(1.7)	0.05	-0.35(2)	-0.4	0.3(1.7)	0.1
FB2	41	1(2.9)	0.7	1.4(4.8)	1	1(2.5)	0.7
FB3	42	-0.3(1.8)	-0.3	-1(2.2)	-0.9	-0.3(1.7)	-0.1
FB4	43	0.001(2.1)	-0.01	0.001(3.7)	0.6	0.1(1.8)	-0.05
FB5	44	1(2.9)	0.7	1.4(4.8)	1	1(2.5)	0.7

Table 2.5: Summary of the values for the timepoint variables related to the fluids measurements. The abbreviation of the variable, its x_i subscript, the mean, standard deviation and median of the variable within the study group, patients who got diuretics and within the patients who did not get diuretics are reported.

Chapter 3

Propensity Analysis

3.1 Introduction

Randomized controlled trials (RCTs) typically compare balance in baseline covariates between treated and untreated subjects. They are a type of scientific experiment, a form of clinical trial, most commonly used in testing the safety (or more specifically, information about adverse drug reactions and adverse effects of other treatments) and efficacy or effectiveness of healthcare services (such as medicine or nursing) or health technologies (such as pharmaceuticals, medical devices or surgery).

An *observational study* is an empirical investigation of treatments and of the effects they cause, but it differs from a randomized controlled trial in the fact that the investigators can't control the assignment of the treatments to the subjects. Observational studies are, by nature, non-randomized and retrospective. Therefore, there is no reason to assume that baseline covariates will be balanced in expectation between treated and untreated subjects. Indeed, treated subjects tend to differ systematically from untreated subjects. Consider the comparison between two heart surgeons, both of them have completed 100 surgeries. The first one had 10 deaths, while the second 5. Apparently the second one would seem to be the best, but how can the two surgeons be compared if the patients of the first one were older and had a higher risk compared to those of the second surgeon?

Under the example scenario presented above, it is important and necessary to seek group of patients under both the doctors that are alike in the statistical sense. This could be achieved by forming sub-groups of patients and then assessing balance in the covariates among these sub groups. Several authors have proposed methods for assessing balance in observational studies. Recent efforts to address issues of nonrandom assignment, includ-

ing a class of methods known as *Propensity Scoring*, can reduce bias in the estimation of treatment effects when assignment is not random.

Propensity score techniques are useful in this context, that is when there may be important differences in patient characteristics between treated and not treated groups. In fact, this kind of medical analysis aims to show whether the differences in the outcomes are attributable to the differences in the treatments provided to the patients, when sometimes it is infeasible or unethical to assign patients to different treatments.

As shown in Figure 3.1, the propensity score method has gained an increasing interest during the last decade. By counting the publications is shown that the number of papers rose sharply from < 10 in 1997 to > 200 in 2007. Through the propensity score, it is possible to produce

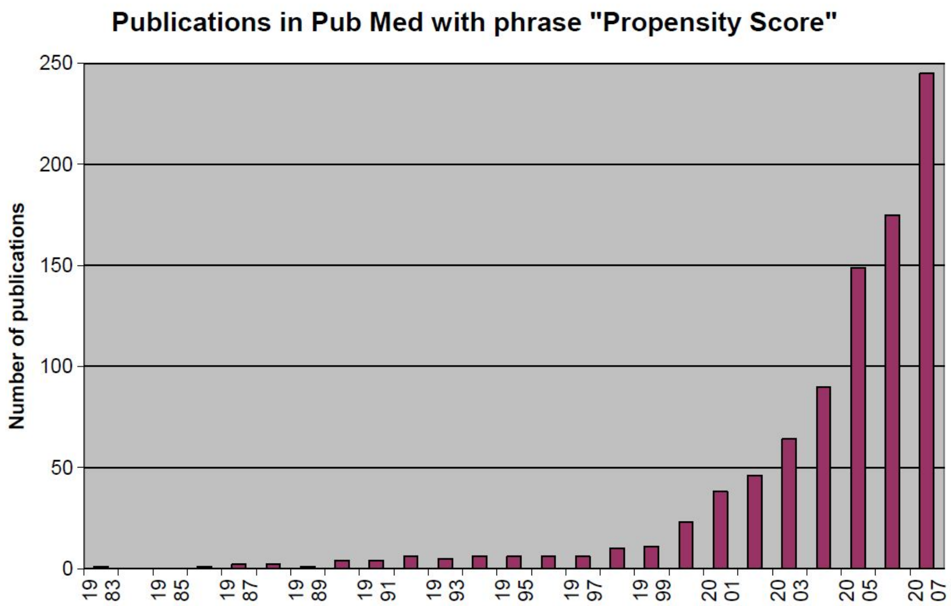


Figure 3.1: The publication regarding the propensity score are increasing in the last three decades.

comparable groups under some nonrandomized conditions. It provides a way to summarize covariate[1] information about treatment selection into a

[1]A covariate is a variable that is possibly predictive of the outcome under study. A covariate may be of direct interest or it may be a confounding or interacting variable. A confounding variable is an extraneous variable in a statistical model that correlates (positively or negatively) with both the dependent variable and the independent variable. An interaction variable may arise when considering the relationship among three or more variables, and describes a situation in which the simultaneous influence of two variables on a third is not additive.

scalar value.

In the next Section the original definition of this technique provided for the first in 1983 time by Rosenbaum and Rubin in [6] will be described.

3.2 Summary of the Dataset

A brief summary of the dataset assembled in Chapter 2 is now provided. There is a total of 1,522 patients in the study group. Out of these 189 recieved diuretics and 1333 did not. In the subsequent sections these patients will be referred to as D^+ and D^- respectively. For each of these patients there are:

(a) 55 variables which will be referred to as covariates.
(b) 1 diuretics decision variable.
(c) 2 outcomes, i.e., 30-day mortality and length of stay in ICU.

3.3 Propensity score model building and balancing

A propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Consider the study group in which are compared two treatments, labeled 1 and 0, denoted by the variable z representing the treatment assignment. Each of the patient is represented by a set of covariates $\mathbf{x} = \{x_1, x_2, ..., x_{55}\}$. The propensity score is the conditional probability that a patient with vector $\mathbf{x}$ of observed covariates will be assigned to treatment 1 given by:

$$e(\mathbf{x}) = Pr(z = 1|\mathbf{x}). \tag{3.1}$$

A systematic approach to build the model and refine it is presented in Rubin and Rosenbaum[6]. The method follows the 4 steps:

Step 1 : Building a propensity model via stepwise logit model.
Step 2 : Stratification and balance assessment.
Step 3 : Refinement of the model.
Step 4 : Decision if the desired balance is achieved or goto Step 2.

An overview of the process is shown in Figure 3.2 on the next page. In the following Subsections details about each step will be provided and the results achieved on the dataset created in Chapter 2 will be presented. The goal of propensity score model building and refining is to find the subgroups (a.k.a

subclassifications) of the patients along the propensity score axis such that within each subclass, patients who are D^+ and D^-, are statistically similar in the covariate space. Statistical similarity is measured by analyzing the difference in each covariate values between the D^+ and D^-. Refinement of the propensity score model is achieved by adding variates or their interaction terms resulting in:

(a) changing of the propensity score values for the patients.
(b) changing the subclass membership of a few patients (if not all).
(c) improving the statistical similarity of covariates within each group.

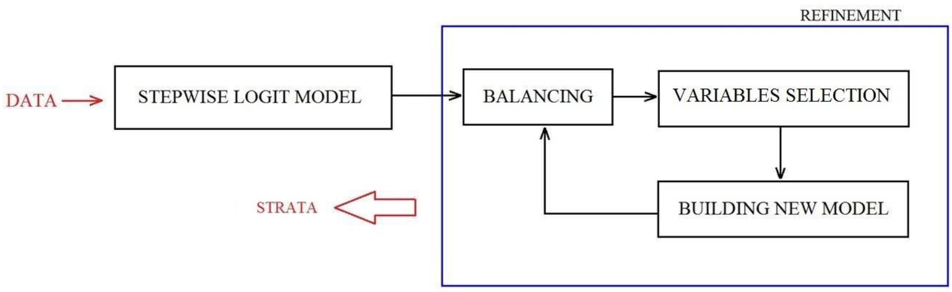


Figure 3.2: According to Rosenbaum and Rubin, the propensity score method is composed by a first phase where a stepwise discriminant analysis is performed on the whole dataset and a second iterative phase in which the achieved balance is evaluated and improved.

3.3.1 Step 1: Building a Propensity Score Model via Stepwise Logit Model

The propensity score is estimated using a logit model (Cox 1970) for

$$e(x) = \frac{e(y)}{1 - e(y)} = \alpha + \beta^T f(x), \tag{3.2}$$

with $y = log[\frac{e(x)}{1-e(x)}]$ and where α and β being parameters and $f(\cdot)$ a specified function determined with the regression model. To build the first propensity score model first the 55 covariates have been provided, to choose from to the stepwise discriminant analysis. The stepwise discriminant analysis method selects a subset of variables $\mathbf{x_s} \in \mathbf{x}$. Then the pariwise interaction terms $x_i \cdot x_j$ where $x_i, x_j \in \mathbf{x_s}$ is provide to the stepwise discriminant analysis method. Each time the parameters for the logit model are estimated via maximum likelihood method[6]. In the study group, the variables (and their pairwise interaction terms) chosen at the end of this step are shown in Table 3.1 on the facing page.

Variable	Detail
x_{11}	Sofa mean of values during the first day
x_{12}	Sofa mean of values during day T1
x_{16}	Elixhauser congestive heart failur
x_{17}	Elixhauser cardiac arrhythmias
x_{40}	Balance average of sums from day 1 to day T1
x_{41}	Balance sum of values during the first day (41)
x_{43}	Balance sum of values during day T2
x_{45}	Use of vasopressors
x_{46}	Mechanical ventilation
x_{47}	Arterial bp average from day 1 to day T1
x_{55}	Arterial bp mean mean of values during day T2
$x_{11} \cdot x_{55}$	-
$x_{40} \cdot x_{43}$	-
$x_{40} \cdot x_{46}$	-
$x_{41} \cdot x_{43}$	-
$x_{43} \cdot x_{43}$	-
$x_{46} \cdot x_{55}$	-

Table 3.1: Covariates and their interactions selected after first step.

3.3.2 Step 2: Stratification and Balance Assessment

The propensity score model built in the previous section provides a score for each patient. Consequently, the patients have been subclassified into 5 groups each group corresponding to a quintile of the distribution of the estimated propensity score. Figure 3.3 illustrates the process of subclassification. It is suggested in Rubin and Rosenbaum that by performing sub-

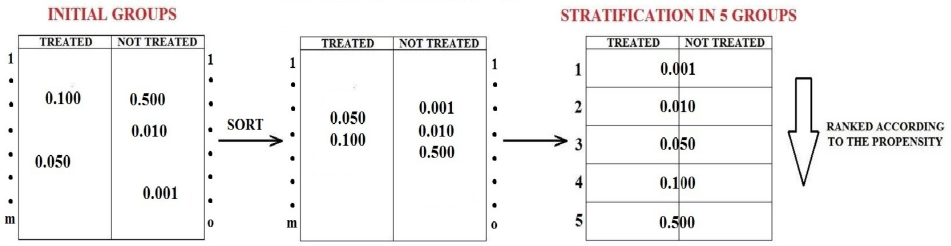


Figure 3.3: The stratification process consists in the ranking of the patients according to their propensity score and then in the division of the whole ranked datasets in 5 quintile. Rosenbaum and Rubin suggest that 5 quintilescan reduce the 90% of the bias in the original dataset.

classification into *quintiles* based on the propensity score, one can largely balance all observed covariates.The balance is achieved **x**, in the sense that within subclasses that are homogeneous in $e(x)$, the distribution of **x** is the same for treated and control (not treated) patients. Note that while all the covariates are not included in the propensity score model, the balance is still sought across all the covariates. In fact this is a key point in the propensity score methodology. At this stage the effectiveness of the subclassification due to this specific propensity score model can be measures by following the method in[6]. The effectiveness is quantified by calculating F-Ratios. The statistical technique, which uses F-ratios, used to assess the balance is briefly explained in Section 3.3.3, but more details will be provided in Appendix B.

3.3.3 Assessing the Balance with Subclasses

To assess balance each of the 55 covariates are subjected to a two-way (2 (treatments) x 5 (subclasses) analysis of variance (ANOVA). In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes T-test to more than two groups. By doing this two F-values for each covariate are calculated. The first one is for treatment vs no treatment interaction. The second one is

for treatment vs subclass interaction. The first value will be called *primary effect* and the second one, *secondary effect*[2].

The achieved balance has been analyzed by comparing a five-number summary (that is minimum, lower quartile, median, upper quartile, maximum) of the 55 F-ratios[3] prior to subclassification with the F-ratios for the primary effect of the treatment and the treatment x subclass interaction in the two-way analysis of variance. The five point summary prior to propensity score modeling and after the first step of propensity score modeling and startification is presented in Figure 3.4 on the following page. Note that the summary statistics are the same for primary and secondary effects initially since the groups have not been stratified yet. It can be observed that after this first step, the F-ratios referring to both primary effects and secondary are decreased significantly, indicating the improvement in the balances of the respective groups subsequent to the propensity score method.

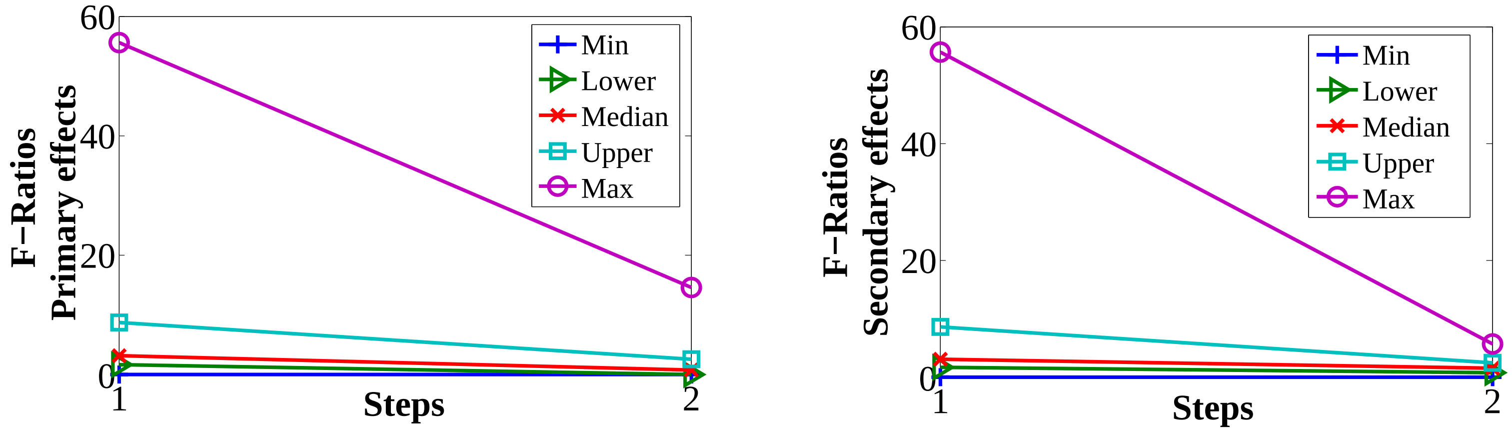
3.3.4 Step 3: Refinement of the Model

In this step, the refinements of the existing logistic model for propensity have been performed to improve the covariates balance further. Covariates with large F-ratios that had previously been excluded from the model were considered for addition to the model. Adding these variables changes the propensity scores for the patients, resulting in reassignment of patients to different quintiles (groups or subclasses). After adding each variable, a logistic models was fitted by maximum likelihood. If the variable produced a lower F-ratio, it was kept. If the variable produced a large F-ratio after inclusion in the model, the square of the variable and cross-products with other variables were instead tried, per the advice of [6].

This refinement process added 11 of 44 variables, that is 25% of them. Figure 3.5 on page 35 shows the improvement made by the inclusion of each variable in the new refined model. It appears that the most important improvements are due to the first few variables which have the biggest F-ratios. Figure 3.6 on page 37 shows the balance achieved in the final refined model, which is considered satisfactory. Note that, if the improvements in the max-

[2]In [6] the first F-ratio hs been called as main effect and the second one as interaction effect. In order to avoid confusion between these definitions and the previous definitions of main effects, being the variables themselves, and interactions effects, representing pairwise products $x_i \cdot x_j$, alternative names have been chosen.

[3]An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fit to a data set, in order to identify the model that best fits the population from which the data were sampled.



a: The F-Ratio statistics after the first step for primary effects.

b: The F-Ratio statistics after the first step for secondary.

Figure 3.4: The F-Ratio statistics after the first step. The values pertaining to NONE refer to the balance on the original dataset, while the values pertaining to ONE refer to balance after the propensity score method.

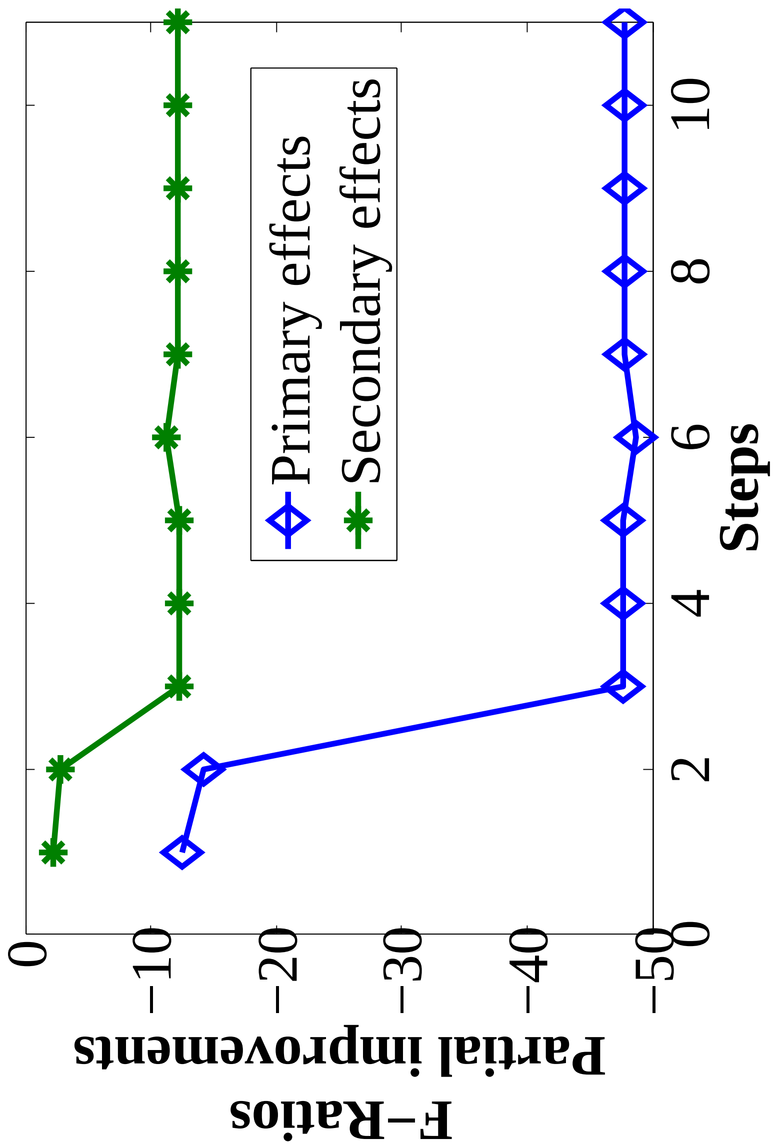


Figure 3.5: The improvement made by the inclusion of each variable in the new refined model.

imum quintile are not considered, there is no substantial improvement in the balance comparing the groups formed after the refinement process.

The complete list of variables included in this final model, of which the first 17 variables were already in the logit model, are in Table 3.2.

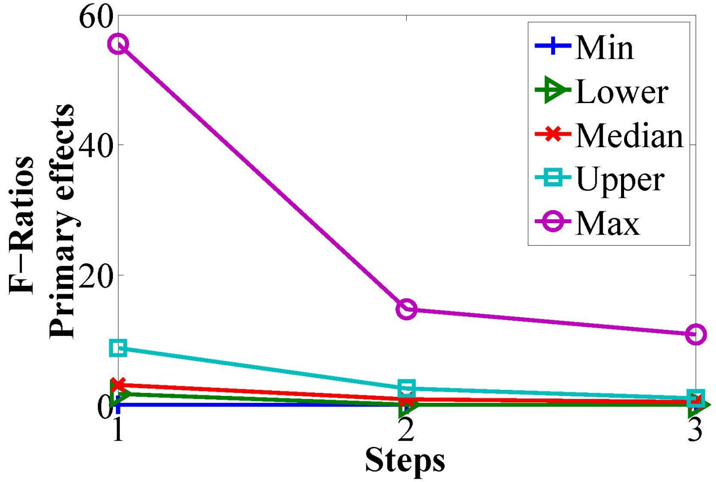
Variable	Detail
x_{11}	Sofa mean of values during the first day
x_{12}	Sofa mean of values during day T1
x_{16}	Elixhauser congestive heart failur
x_{17}	Elixhauser cardiac arrhythmias
x_{40}	Balance average of sums from day 1 to day T1
x_{41}	Balance sum of values during the first day (41)
x_{43}	Balance sum of values during day T2
x_{45}	Use of vasopressors
x_{46}	Mechanical ventilation
x_{47}	Arterial bp average from day 1 to day T1
x_{55}	Arterial bp mean mean of values during day T2
$x_{11} \cdot x_{55}$	-
$x_{40} \cdot x_{43}$	-
$x_{40} \cdot x_{46}$	-
$x_{41} \cdot x_{43}$	-
$x_{43} \cdot x_{43}$	-
$x_{46} \cdot x_{55}$	-
x_{20}	Elixhauser diabetes uncomplicated
x_{32}	Inputs sum of values during day T1
x_{42}	Balance sum of values during day T1
x_8	Saps mean of values during day T2
$x_{12} \cdot x_5$	$x_{12}\cdot$ Saps average of sums from day 1 to day T1

Table 3.2: Covariates and their interactions selected after first step.

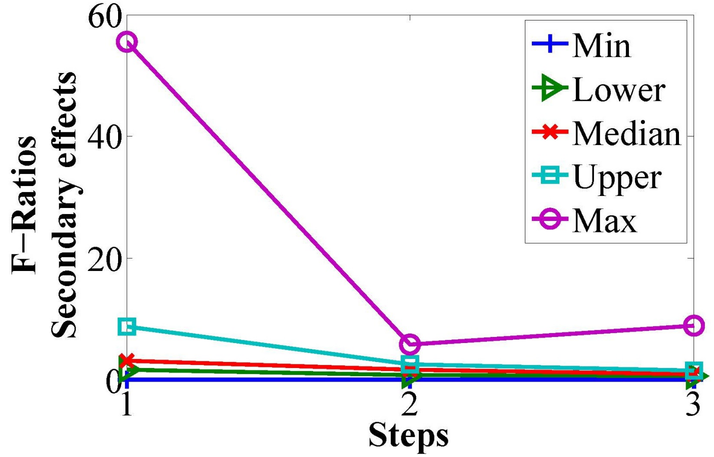
3.3.5 Experts' Covariate Sets

Finally, the analysis was repeated for 2 new variable sets based on advice by clinical experts. The details of this analysis are presented in Appendix C. Note that, if the improvements in the maximum quintile are not considered, there is no substantial improvement in the balance comparing the experts' variable sets to the final refinement-based balance.

Figure 3.7 on page 39 shows the variables which were selected in each



c: The F-Ratio statistics after the first step for primary effects.



d: The F-Ratio statistics after the first step for secondary.

Figure 3.6: The F-statistics on the refined dataset on the primary and secondary effects after the refinement. The values compare the original balance, the balance after the first logit model and the one in the final model.

of 4 variable sets: the set selected in the *automatic* stepwise discrimination process before iterative refinement, the set selected in the *refined* model and the two sets resulting from starting with the experts' variable sets. Five variables were select in all four variable sets: Elixhauser congestive heart failure (x_{16}), Elixhauser cardiac arrhythmias (x_{17}), Fluids balance sum of values during the first day (x_{41}), Use of vasopressors (x_{45}) and Mechanical ventilation (x_{46}).

3.3.6 Estimating the Average of Treatment Effects

In the Rosenbaum and Rubin study,[6], the groups defined by the propensity analysis, which are now homogeneous, are directly compared with respect to basic statistics on mortality or on other predetermined outcomes.

Austin warns that when using propensity score methods involving pair matching between patients, the matched nature of the pairs should be considered in the analysis of the outcomes [7]. But in the creation of the quintile, stratification is performed prior to treatment assignment (as in randomized controlled trials) and this implies that there is no reason subjects within a stratum are more similar then randomly selected ones.

Considering this, it is possibile using the stratification approach to estimate the treatment effects of the treatment just by direclty comparing the treated and not treated groups.

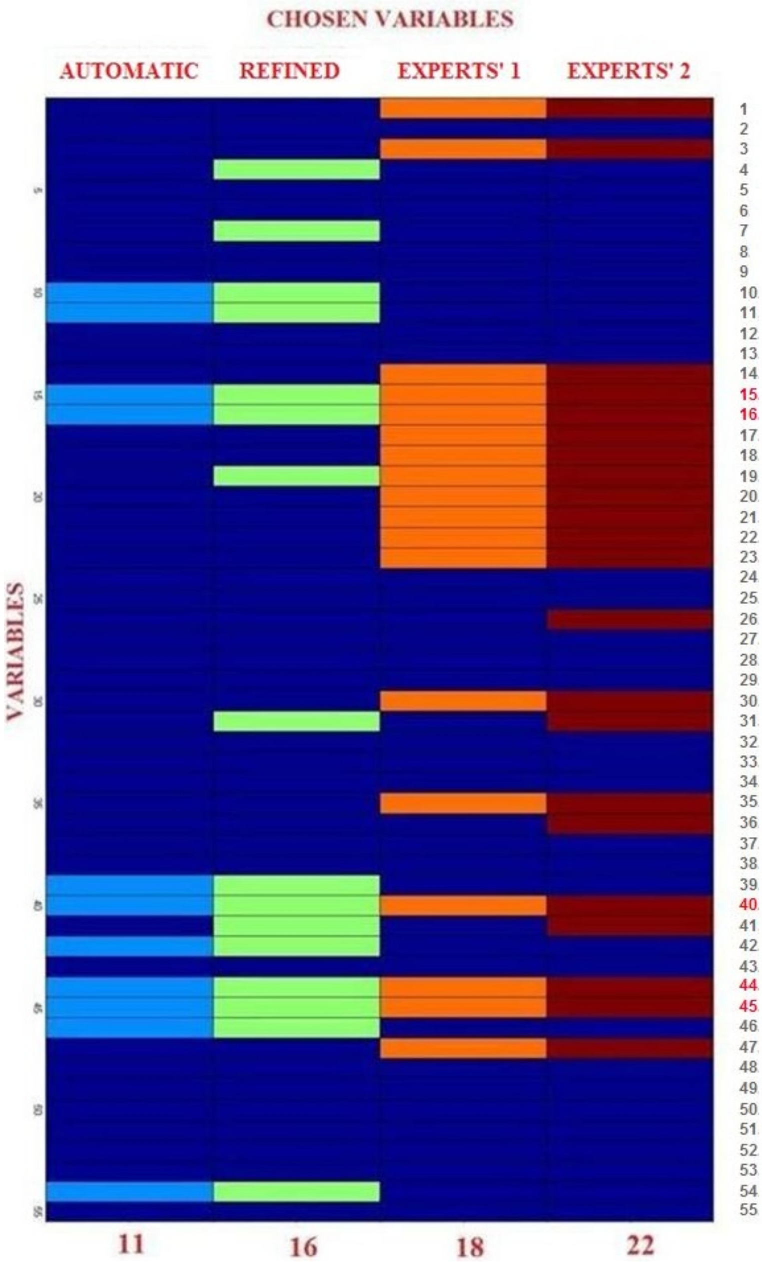


Figure 3.7: Five variables were chosen from all the four variable sets: Elixhauser congestive heart failure (x_{16}), Elixhauser cardiac arrhythmias (x_{17}), Fluids balance sum of values during the first day (x_{41}), Use of vasopressors (x_{45}) and Mechanical ventilation (x_{46}).

3.4 Stratification Results

Table 3.3 on the next page shows the quintiles, their propensity score ranges and their split between D^+ and D^- in terms of number of patients, mortality rate and mean length of ICU stay for the propensity score model generated automatically by stepwise discrimination, before refinement. The stepwise logit model in this case maximizes the accuracy of the model, however, in fact, the first two quintiles are very unbalanced in the number of treated and untreated patients and cannot considered useful for analysis. Quintiles 3 and 4 are less unbalanced even if the number of untreated patients is still a lot bigger then the treated ones. In quintile5 there are balanced numbers.

From the results, it appears that typically the treated patients spend more time in the ICU. The mortality however seems to be more or less the same, except for group 3 where diuretics seem to slightly improve the chances of survival.

Table 3.4 on page 42 shows the quintiles, their propensity score ranges and their split between D^+ and D^- in terms of number of patients, mortality rate and mean length of ICU stay for the propensity score model which was subsequently iteratively refined. Even though the variable set went through the refinement, quintiles 1 and 2 are still very unbalanced. Quintiles 3 and 4 are less unbalanced, while Quintile 5 is balanced. Except for quintile 5, also in this dataset it seems that patients who got diuretics have a slightly better chance of survival.

Quintile 1 $PS \in [0.00; 0.00]$	Diuretics given	Diuretics not given
Number of patients	2	302
Deaths	0%	43%
Mean length of stay	40 days	3.7 days
Quintile 2 $PS \in [0.00; 0.02]$	Diuretics given	Diuretics not given
Number of patients	5	299
Deaths	40%	32%
Mean length of stay	28 days	4.1 days
Quintile 3 $PS \in [0.02; 0.06]$	Diuretics given	Diuretics not given
Number of patients	20	284
Deaths	15%	30%
Mean length of stay	13.3 days	6 days
Quintile 4 $PS \in [0.07; 0.17]$	Diuretics given	Diuretics not given
Number of patients	27	277
Deaths	37%	40%
Mean length of stay	13.5 days	9.5 days
Quintile 5 $PS \in [0.17; 1.00]$	Diuretics given	Diuretics not given
Number of patients	134	170
Deaths	41%	44%
Mean length of stay	15 days	10 days

Table 3.3: Results on the Automatic generation dataset.

Quintile 1 $PS \in [0.00; 0.01]$	Diuretics given	Diuretics not given
Number of patients	4	300
Deaths	25%	30%
Mean length of stay	46.7 days	1.3 days
Quintile 2 $PS \in [0.01; 0.04]$	Diuretics given	Diuretics not given
Number of patients	4	300
Deaths	25%	33%
Mean length of stay	7 days	5.2 days
Quintile 3 $PS \in [0.04; 0.08]$	Diuretics given	Diuretics not given
Number of patients	25	279
Deaths	24%	38%
Mean length of stay	13.8 days	7.4 days
Quintile 4 $PS \in [0.08; 0.18]$	Diuretics given	Diuretics not given
Number of patients	27	277
Deaths	29%	46%
Mean length of stay	10.7 days	8.8 days
Quintile 5 $PS \in [0.18; 0.99]$	Diuretics given	Diuretics not given
Number of patients	127	177
Deaths	45%	42%
Mean length of stay	15.7 days	10.5 days

Table 3.4: Results on the Automatic generation dataset after the refinement process.

3.4.1 Comparison between the Quintiles

Now a comparison between the quintilesfor the Refined dataset will be discussed. In Figure 3.8 shows a parallel between the number of patients for the 5 quintile. The numbers seem to be consistent as going from quintile 1 to 5, there is an increasing number of patients who received diuretics as the propensity of getting them is increasing. Figure 3.9 on the following page

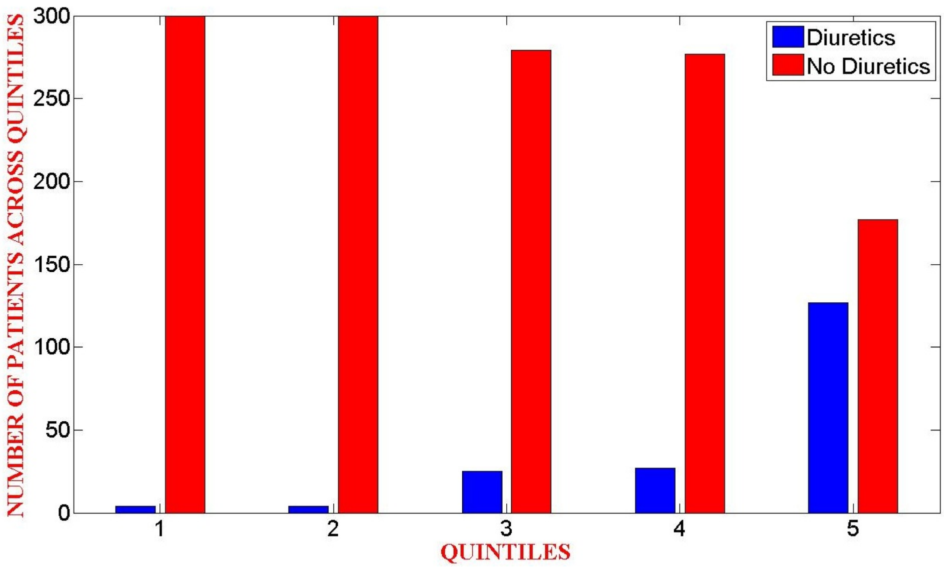


Figure 3.8: Number of patients in the refined dataset, across the quintiles.

show a comparison between the mortality rate in the 5 quintile. In the first 4 quintilethere is a slightly better chance of survival by the administration of diuretics, while for quintile5 the chances are similar. Figure 3.10 on the next page show a comparison between the length of stay in ICU in the 5 quintile. The results seem consistent except for the first quintile, where probabily there are outliers or noisy values for the patients who got diuretics. It seems that the patients treated with diuretics have a longer stay in ICU.

3.4.2 Comments on the Results

In both the 2 datasets, mortality rates are usually similar between treated and untreated patients and looking at those values, it seems that the treated patients have a slightly better chance of survival. But from this analysis is not easy to determine in which cases the diuretics should be given, it can just be concluded that, in case of indecision, they may be of benefit for the patient. Instead, in all the four datasets, seems that the treated patients

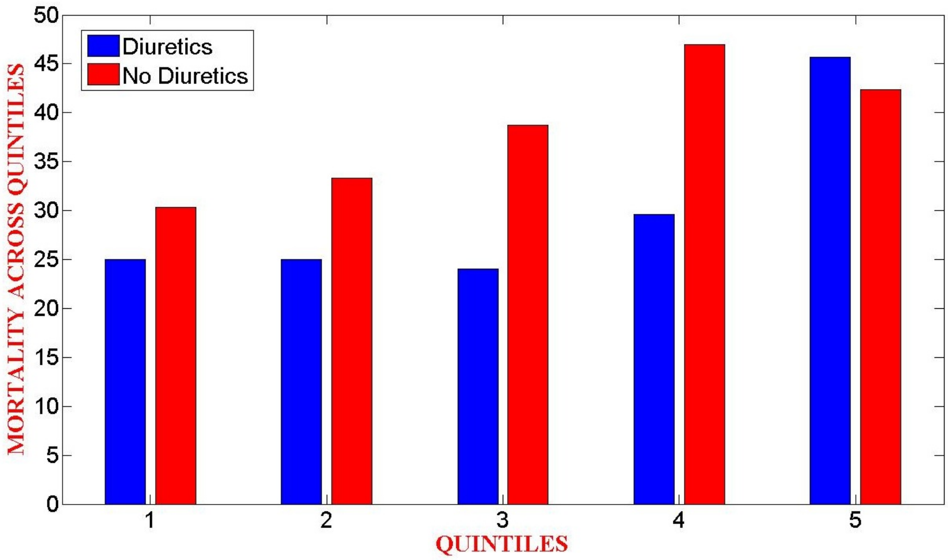


Figure 3.9: Death balance in the 5 quintiles.

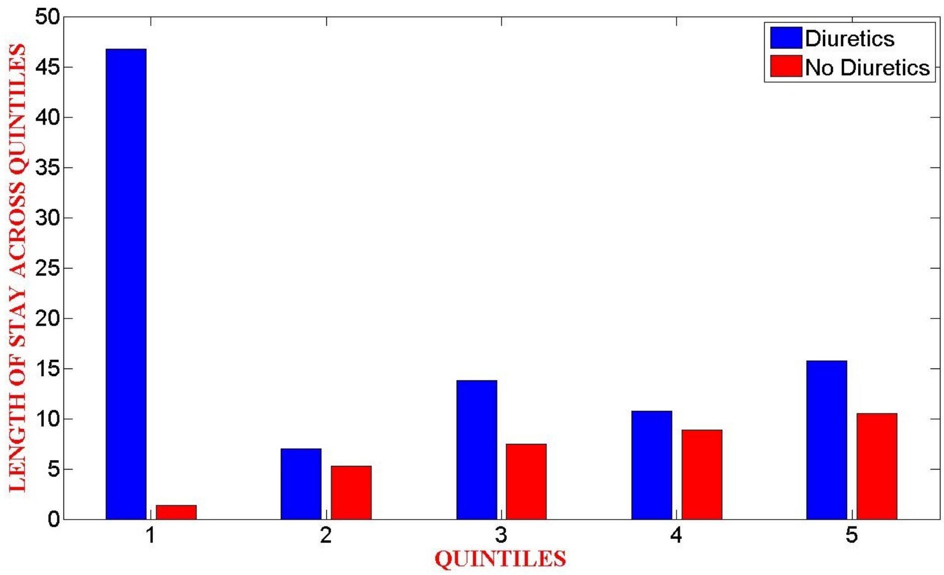


Figure 3.10: Length of stay in ICU balance in the 5 quintiles.

Automatic Dataset	Deaths	Length of Stay
Diuretics given	31%	13.9 days
Diuretics not given	38%	8.5 days
Automatic Refined Dataset	Deaths	Length of Stay
Diuretics given	32.6%	13.4 days
Diuretics not given	42%	8.9 days

Table 3.5: Results on quintiles 3, 4 and 5. The patients to whom diuretics were given seems to have a better chance of survival, while they have a longer stay in the ICU.

stay more time in the ICU.

In Table 3.5 are shown the averages of the outcomes in the 2 models. The averages are made considering only quintile 3, 4 and 5 as quintile 1 and 2 have too few patients to whom diuretics were administrated. The patients to whom diuretics were given seem to have a better chance of survival, while they have a longer stay in the ICU. In Table 3.6 on the next page is shown the same analysis performed using all 5 quintile. Comparing the two tables, it is possibile to see that the values for the length of stay for the patients who got diuretics of the first two datasets seem to be outliers: in these groups the patients who actually got the drugs were too few to have a reliable result.

It is possibile to see another reason to use only quintile3, 4 and 5 by noticing that in the results of the analysis performed on all the 5 quintileboth the mortality rates and the lengths of stay in the ICU in all the four datasets are better, in fact the chance of survival is higher and the lenght of stay in the ICU shorter. This result is consistent with what was said by doctors involved in the work: clinicians are usually against giving diuretics because in general they are harmful drugs and if they give them, the length of stay in ICU of the patient lengthens. In the tables, going from the quintile1 to 5, the propensity for diuretics of being prescribed increases, and this goes along with the worsening of the conditions of the patients: by using quintiles3, 4 and 5 the analysis is capturing the chances of survival of the patients on the border line, that is the patients whose conditions leave to the doctors the judgment whether giving or not diuretics. Furthermore, have to be said that in group 5 there are the sicker patients: in this groups seems not to make any difference if diuretics were given or not. This could mean that quintile3 and 4 are the most relevant for the analysis, as in these groups the fact that diuretics were given or not seems to make the difference in the survival of the patients. In Table 3.7 on the following page are shown the results for these two groups.

Automatic Dataset	Deaths	Length of Stay
Diuretics given	26.6%	21.9 days
Diuretics not given	37.8%	6.6 days
Automatic Refined Dataset	Deaths	Length of Stay
Diuretics given	29.6%	18.7 days
Diuretics not given	37.8%	6.6 days

Table 3.6: Results on all the 5 quintiles. The mortality rates are decreasing as a bigger propensity of getting diuretics is related the a worse condition of the patient.

Automatic Dataset	Deaths	Length of Stay
Diuretics given	26%	13.4 days
Diuretics not given	35%	7.7 days
Automatic Refined Dataset	Deaths	Length of Stay
Diuretics given	26.5%	12.2 days
Diuretics not given	42%	8.1 days

Table 3.7: Results on quintiles 3 and 4. The differences of the chances of survivals between treated and untreated patients are wider in this case.

The percentage of death and length of stay in the ICU have also been calculated on the whole (without the propensity method) dataset. The results on the original dataset are, comparing the patients to whom diuretics were given to the ones who didn't got diuretics, the 32.8% vs 37.8% percentage of death and 15.1 vs 6.3 days in the ICU. These results are shown in Table 3.8. As predicted, here the differences of the mortality rates are slightly narrowing because is not considered the fact that usually the patients to whom diuretics are administrated are sicker, and have a worst chance of survival.

3.4.3 Conclusions of the Propensity Analysis on the Diuretics Problem

In conclusion from this analysis, only comparing the outcomes within the quintile, seems that patients who received diuretics have a slightly better

Original Dataset	Deaths	Length of Stay
Diuretics given	32.8%	15.1 days
Diuretics not given	37.8%	6.3 days

Table 3.8: Results on the original dataset, without propensity analysis and stratification.

chance of survival, while they stay longer in the ICU even if it is not clear how statistically relevant these results are and when exactly diuretics should be given.

The refined dataset seems to be the best one, and on it have been performed a series of statistic tests to decide if the results should be considered statistically significant.

The Chi-squared test[4] has been performed to compare mortality between the 5 quintileand the T-test[5] has been used for length of stay.

From the tests appears that the results for mortality are not statistically significant in all the 5 quintile, while the ones for length of stay are significant for quintile1, 3 and 5.

This confirms that is not clear if diuretics are harmful or not: in fact, even if in the 2 datasets, when comparing the percentages of death between the quintilebetween the treated and untreated patients the chances of survival are increasing of 7% or 8%, these differences are not statistically significant and they may be randomly happened. Instead it is easier to deduce that the length of stay in the ICU increases for the patients who got diuretics, according to the average results between 5 and 6 days.

From this analysis can be concluded that, in general, diuretics should not be given as they do not seem to make difference in the chances of survival of the patients, while by giving them the length of stay in ICU is lengthened.

This analysis can not provide any information on the conditions when diuretics should be given and this will be the object of Chapter 4.

[4]A Chi-squared test is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true, or any in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough. It can be used for dichotomous variables, as mortality.

[5]A T-test is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported. It is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's t distribution. It can be used for continuous variables, as length of stay.

Chapter 4

Outcome Analysis

4.1 Introduction

In this Chapter, Step 3 of the analysis, will be described. While the propensity score analysis of Chapter 3 provides balanced quintiles of patients with respect to propensity of the administration of diuretics, it does not account for confounding factors which might also affect mortality and length of stay. An obvious confounding factor is a patient's health condition, i.e. underlying illness. In this Chapter the question that have been answered is: **Does the administration of diuretics have a statistically significant effect on mortality or length of stay? If so, to what extent?**

Therefore, is first described a modeling methodology for:

A. Determining if, when health condition is taken into account, the administration of diuretics has a significant effect on outcome (mortality or length of stay).

B. Determining, if the administration of diuretics has no significant effect in (A), whether the cross-variable interaction of the administration of diuretics and health condition, has a significant effect on outcome (mortality or length of stay).

C. Given (B), i.e. that the administration of diuretics has a significant effect on outcome, determining if the administration of diuretics crossed with health condition has a significant effect on outcome when the study group is adjusted according to health condition.

Each determination involves the regression of a model and statistical determination of effect. There is 1 step in the methodology corresponding to each of the determinations:

Step A. Use the study group to regress propensity score, health condition and diureticsDecision as independent variables for the outcome mortality using a logistic regression. For the length of stay outcome, use generalized linear regression. These models will be labelled by ModelA and append to the model label either 'Mortality' or 'LOS' for length of stay, e.g. ModelA.Mortality and ModelA.LOS .

Set up a null hypothesis that diureticsDecision has no significant effect on outcome. Examine the p-value of diuretics decision variable. If the p-value < 0.05, accept the null hypothesis. If the null hypothesis is rejected, revisit the balanced propensity quintiles and consider the result of a chi-squared test for significance of difference in outcome between patients the administration of diuretics and not with the administration of diuretics to be conclusive.

For more information on statistical hypothesis tests and use of p-value see Appendix D.

Step B. If the effect of the administration of diuretics is NOT significant, use the study group to regress the same variables as Step A while adding a new variable SAPS-$t0$_diureticsDecision. Such a model will be labelled as ModelB. Set up a null hypothesis that the new cross-variable SAPS-$t0$_diureticsDecision has no significant effect on outcome. Examine the p-value of this variable. If the p-value < 0.05, reject the null hypothesis and proceed to Step C.

Step C. Divide the study group into 2 subsets according to health condition by using SAPS-T_0's median value as a threshold. Repeat Step A with each subset and evaluate the null hypothesis that diureticsDecision has no significant effect on outcome, in the cases of a subset of sick and another of less sick patients.

In the following sections the modeling methodology will be demonstrated.

4.2 Confounding Factors

A confounding factor in a study is a variable which is related to one or more of the variables defined in the study. A confounding factor may mask an actual association or falsely demonstrate an apparent association between the study variables where no real association between them exists. If confounding factors are not measured and considered, bias may result in the conclusion of the study[8].

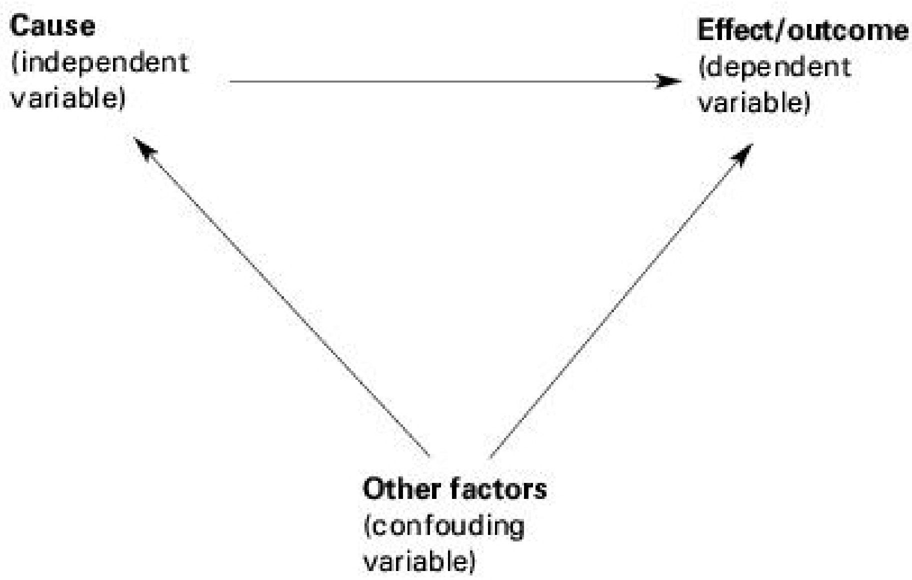


Figure 4.1: A confounding factor in a study is a variable which is related to one or more of the variables defined in a study. A confounding factor may mask an actual association or falsely demonstrate an apparent association between the study variables where no real association between them exists. If confounding factors are not measured and considered, bias may result in the conclusion of the study.

ModelA		Mortality p-value	Mortality $\beta_{i,1}$	LOS p-value	LOS $\beta_{i,2}$
x_1	**DiureticsDecision**	**0.075**	**-0.189**	< 0.001	**2.626**
x_2	Age	< 0.001	0.023	< 0.001	-0.078
x_3	Gender	0.410	0.048	0.753	-0.092
x_5	SAPS-T_0	< 0.001	0.053	0.004	0.202
x_{10}	SOFA-T_0	< 0.001	0.125	0.649	-0.039
x_{15}	Elixhauser Score	0.095	0.058	0.019	-0.409
V_1	Propensity Score	0.478	-0.289	< 0.001	11.795

Table 4.1: Effects of variables in ModelA.*Mortality and* ModelA.*LOS. For both mortality and length of stay outcomes, health condition variables x_2, x_3, and x_4 (Age, SAPS-T_0, and SOFA-T_0) have statistically significant effects and thus are highlighted with yellow cells. The Administration of Diuretics instead appears to have a significant effect only for length of stay. The positive β coefficient sign implies length of stay increases when diureticsDecision is true.*

4.3 Step A, ModelA: health condition and propensity adjustment.

The purpose of Step A is to determine, when health condition is taken into account, whether the administration of diuretics has a significant effect on outcome (mortality or length of stay). For the regression of both ModelA.Mortality and ModelA.LOS , the patient's diureticsDecision as x_1 and his/her propensity score (as calculated in Chapter 2) as V_1 have been included. To express health condition the following independent variables x_2, x_3, x_5, x_{10} and x_{15} have been chosen: Age, Gender, SAPS-T_0, SOFA-T_0, Elixhauser Score. Table 4.1 shows the p-value and beta coefficient analyses for both outcomes. With respect to mortality, health condition variables x_2, x_3, and x_10 (Age, SAPS-T_0, and SOFA-T_0) have statistically significant effects. The diureticsDecision does not have a statistically significant effect on mortality. With respect to length of stay, illness variables x_2, x_4, x_{15} (Age, SAPS-T_0, and Elixhauser Score) have statistically significant effects, as does propensity score. Importantly, and in contrast to mortality, the null hypothesis that the effect of diureticsDecision is not significant on length of stay, is rejected (p-value < 0.001). For length of stay outcomes, these findings imply health condition is not a confounding factor and diureticsDecision is independently significant in its effect on Length of Stay. One can go back to each propensity quintile and, where there is sufficient data, examine the T-test of the difference between the length of stay outcome distribution

for patients with the administration of diuretics and those without. In this case, the test indicates for quintile 1, quintile 3 and quintile 5 a significant difference, leading to the conclusion, qualified for this study group, that the administration of diuretics increases a patient's length of stay in the ICU.

4.4 Step B

For mortality outcome, the diureticsDecision is *NOT* independently significant in its effect. However, it may be there is an interactive effect of diureticsDecision with health condition that is more than random. Therefore, what happens in modeling when the cross-interaction variable SAPS-$t0$_diureticsDecision, $x_1 \cdot x_5$, is included will now be examined. Table 4.2 on page 56 columns 2 and 3 show the p-value and Beta coefficient analyses for this mortality outcome logistic regression. The null hypothesis that SAPS-$t0$_diureticsDecision has no significant cross-dependent effect on mortality is rejected given the p-value = 0.013. Therefore, Step C have been performed and two models generated: ModelC.LessSick and ModelC.Sicker using two subsets divided by relative SAPS-T_0 median within the study group.

4.5 Step C, ModelC: Health condition Split and New Adjustment Models

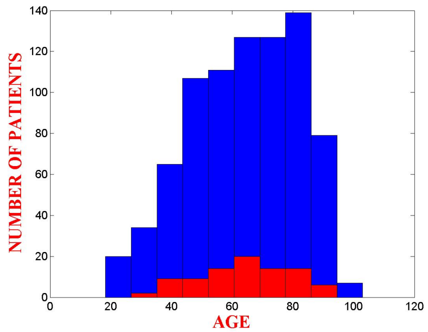
4.5.1 Splitting the Study Group by Health condition

The 2 health condition subsets are divided across the median SAPS-T_0 score of 17. The less sick subset is composed of 816 patients and the sicker subset has 706 patients. Descriptive statistics of the two subsets in terms of Age, SAPS-T_1 and SOFA-T_1 are provided in Figures 4.2 on the following page for the less sick subset and Figures 4.3 on page 55 for the sicker one. Other descriptive statistics of the subsets are provided in Appendix C.

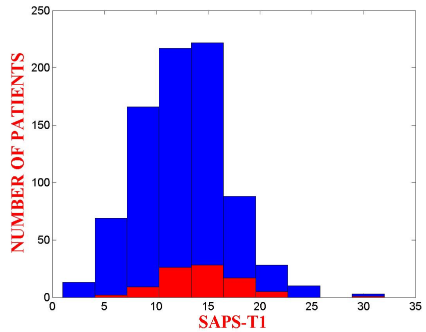
As predicted, all the clinical values for the sicker group are generally worst.

4.5.2 ModelC.LessSick and ModelC.Sicker: New Adjustment Models

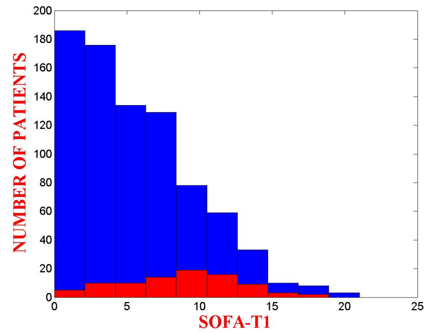
Table 4.2 on page 56 columns 4 and 5 show the p-value and Beta coefficient analyses for the mortality outcome regression on the less sick subset. The null hypothesis that diureticsDecision has a no significant cross-dependent effect on mortality in the less sick subset is rejected. Columns 6 and 7



a: *Age is centered on 65 years.*

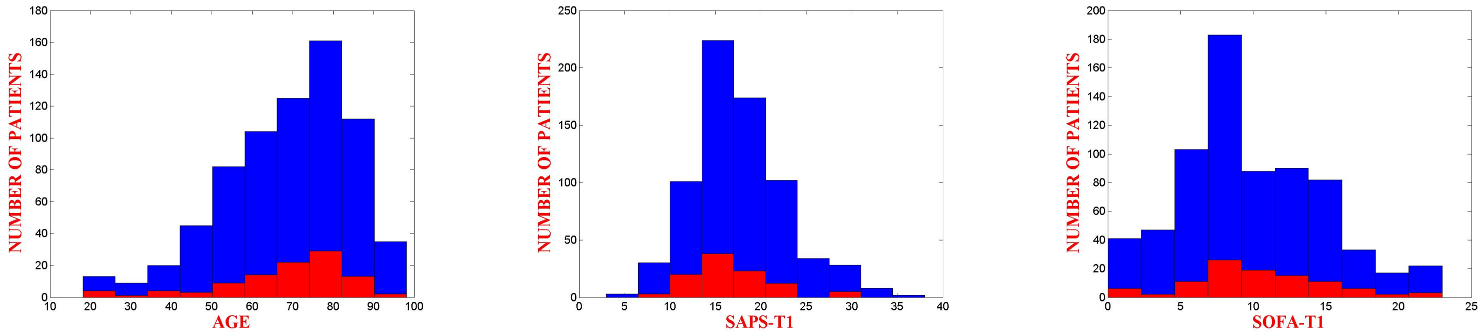


b: *SAPS-T_1 is centered on 13.*



c: *SOFA-T_1 is centered on 5.*

Figure 4.2: Histogram of Age, SAPS-T_1 and SOFA-T_1 in the subsets divided by health condition for ModelC.LessSick. *In red the values for the D^+ patients only.*



d: Age is centered on 71 years. *e: SAPS-T_1 is centered on 17.* *f: SOFA-T_1 is centered on 9.*

Figure 4.3: Histogram of Age, SAPS-T_1 and SOFA-T_1 in the subsets divided by health condition for ModelC.Sicker*. In red the values for the D^+ patients only.*

Var	Model B p-value	$\beta_{i,0}$	Model C Less Sick	$\beta_{i,1}$	Model C Sicker	$\beta_{i,2}$
x_1	0.069	0.602	0.004	1.842	0.531	-0.50
x_2	< 0.001	0.023	< 0.001	0.0234	< 0.001	0.025
x_3	0.347	0.054	0.445	0.064	0.760	0.025
x_5	0.278	0.021	0.270	-0.054	0.244	0.043
x_{10}	< 0.001	0.123	< 0.001	0.094	< 0.001	0.141
x_{15}	0.100	0.057	0.645	0.023	0.067	0.092
V_1	0.576	-0.224	0.093	1.036	0.038	-1.148
$x_1 \cdot x_5$	0.013	-0.043	0.001	-0.145	0.709	0.013

Table 4.2: ModelBAnalysis (columns 2 and 3) indicates that the cross product variable SAPS-t0_diureticsDecision ($x_1 \cdot x_5$) has a significant effect on mortality. ModelCanalysis (columns 4 and 5 for the less sick subset, and columns 6 and 7 for the sicker subset. In the less sick subset, diureticsDecision is a significant independent variable effect, whereas in the sicker subset, it is not (red font).

show the p-value and Beta coefficient analyses for the mortality outcome regression on the sick subset. The null hypothesis that diureticsDecision has a no significant cross-dependent effect on mortality in the sicker subset is *NOT* rejected.

4.5.3 Stratification Analysis with Adjustment for Confounding Factor of Health condition

Returning to quintile analysis: the study group has been divided by SAPS-T_0 median threshold into 2 groups.

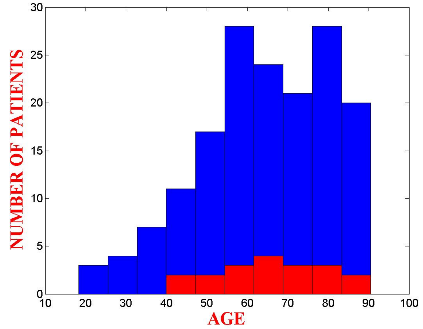
Descriptive statistics of quintiles 4 and 5 in terms of Age, SAPS-T_1 and SOFA-T_1 is provided in Figures 4.4 on page 58 for the less sick subset and Figures 4.5 on page 59 for the sicker one.

All patients of a group are ranked by propensity score and divide the ranked group into 5 quintiles of equal size. In a quintile, in each health condition subset, the mortality rate for those patients with the administration of diuretics and those which did not were compared. In this case the null hypothesis is that the outcomes come from the same distribution. To test the null hypothesis, the Chi-Squared test has been used.

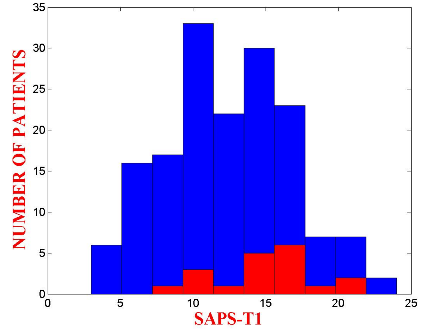
The health condition adjusted stratification analysis is summarized in Table 4.3 on page 60. In the less sick subset of quintile 4 ($PS \in [0.06; 0.15]$) mortality rate is significantly less for the patients with the administration of diuretics compared to those without. It is not significantly different for quin-

tile 5 ($PS \in [0.15; 0.99]$). The mortality rate is not significant in quintiles 1, 2 and 3 either.

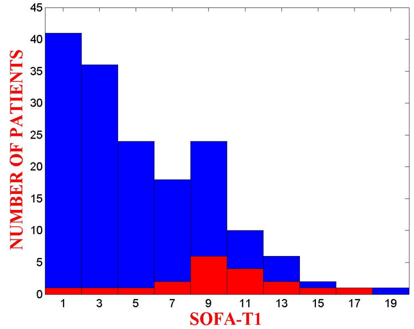
For the sicker group, see Table 4.4 on page 61, mortality rate is not significantly different for quintiles 1 to 5.



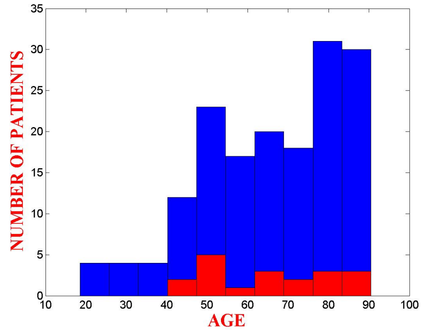
g: Age is centered on 65 years.



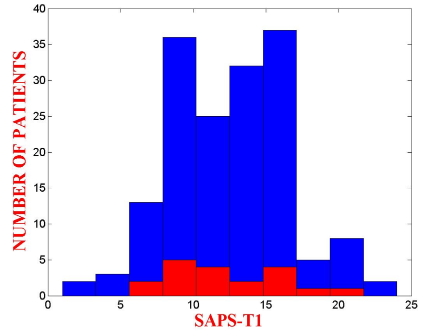
h: SAPS-T_1 is centered on 12.



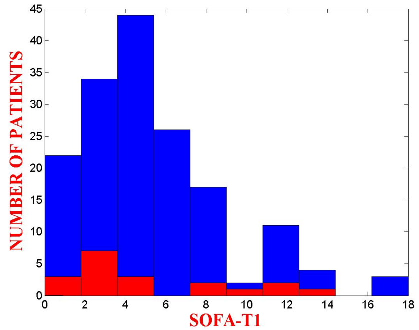
i: SOFA-T_1 is centered on 5.



j: Age is centered on 72 years.

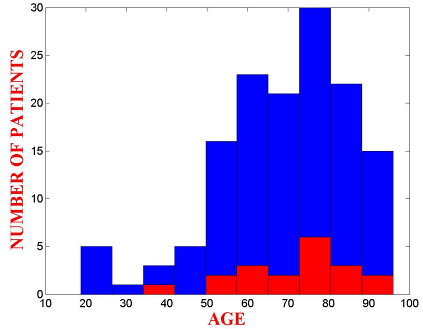


k: SAPS-T_1 is centered on 17.

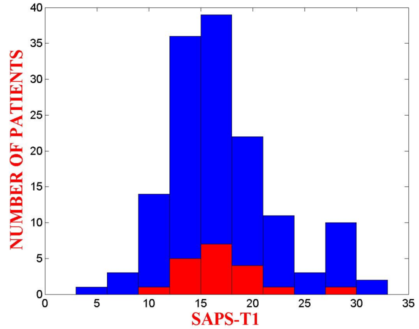


l: SOFA-T_1 is centered on 10.

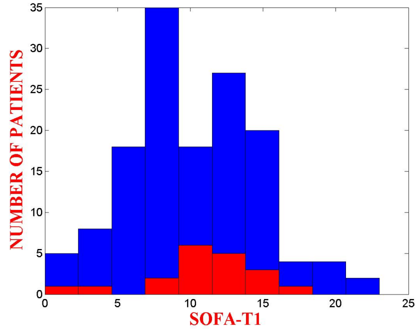
Figure 4.4: Histogram of Age and SAPS-T_0 in quintiles 4 and 5 formed by SAPS-T_0 median for ModelC.LessSick. *In red the values for the D^+ patients only.*



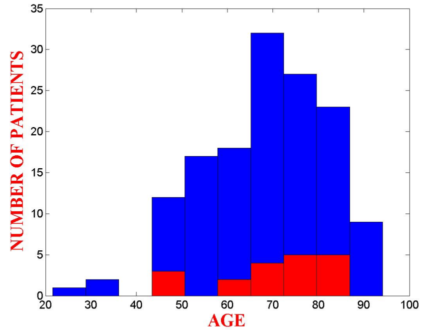
m: Age is centered on 68 years.



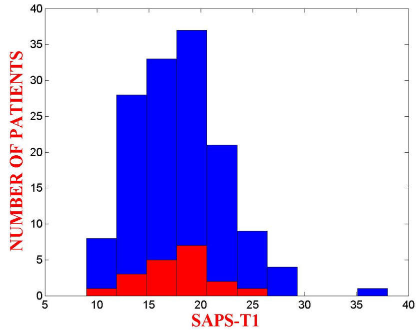
n: SAPS-T_1 is centered on 13.



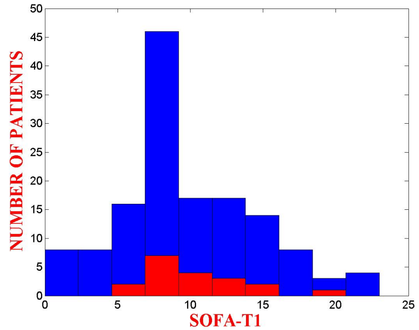
o: SOFA-T_1 is centered on 4.



p: Age is centered on 69 years.



q: SAPS-T_1 is centered on 18.



r: SOFA-T_1 is centered on 9.

Figure 4.5: Histogram of Age and SAPS-T_0 in quintiles 4 and 5 formed by SAPS-T_0 median for ModelC.Sicker. *In red the values for the D^+ patients only.*

quintile 1 $PS \in [0.00; 0.01]$	D^+	D^-
Number of patients	0	163
Deaths	0%	14%
quintile 2 $PS \in [0.01; 0.02]$	D^+	D^-
Number of patients	4	159
Deaths	25%	24%
quintile 3 $PS \in [0.02; 0.06]$	D^+	D^-
Number of patients	6	157
Deaths	33%	24%
quintile 4 $PS \in [0.06; 0.15]$	D^+	D^-
Number of patients	21	142
Deaths	14%	28%
quintile 5 $PS \in [0.15; 0.99]$	D^+	D^-
Number of patients	57	106
Deaths	28%	40%

Table 4.3: Mortality outcomes after propensity and less sick stratification. The difference in mortality is statistically significant (null hypothesis of Chi-Squared test is rejected) in quintile 4 only. The difference in mortality is not statistically significant (null hypothesis of Chi-Squared test is not rejected) in quintiles 1, 2, 3 and 5.

quintile 1 $PS \in [0.00; 0.02]$	D^+	D^-
Number of patients	4	137
Deaths	25%	61%
quintile 2 $PS \in [0.02; 0.05]$	D^+	D^-
Number of patients	4	137
Deaths	25%	46%
quintile 3 $PS \in [0.05; 0.1]$	D^+	D^-
Number of patients	9	132
Deaths	22%	53%
quintile 4 $PS \in [0.1; 0.21]$	D^+	D^-
Number of patients	15	126
Deaths	53%	54%
quintile 5 $PS \in [0.21; 0.99]$	D^+	D^-
Number of patients	68	73
Deaths	55%	45%

Table 4.4: Mortality outcomes after propensity and sicker stratification. The difference in mortality is not statistically significant (null hypothesis of Chi-Squared test is not rejected) in quintiles 1, 2, 3, 4 and 5.

Chapter 5

Machine Learning with GP Analysis

5.1 Introduction

In this Chapter the analysis performed by using GP techniques will be discussed. A description of Machine Learning and Genetic Programming techniques is available in Appendix E. This analysis used GPLAB, A Genetic Programming Toolbox for MATLAB produced by Sara Silva[1]. GPLAB is a genetic programming toolbox for MATLAB and its architecture follows a highly modular and parameterized structure. For a description of the toolbox see[9].

GP was used to classify the study group on mortality and to evolve symbolic regression to predict length of stay. For this analysis the 8 variables in Table 5.1 on the next page were used.

The presented results are preliminary and need further study to tune the GP method properly. This should be considered an initial exploration.

The goal of this analysis was the use of GP-based machine learning (ML) for predictive outcome modeling with the diuretics study as startup demostration context. The envisioned approach to helping a new patient, is to:

A: identify the cluster were the new patient is placed.

B: push each new patient's variables into the identified cluster model.

The used approach is divided into 2 steps:

[1] Sara Silva is currently, Summer 2012, senior researcher of the KDBIO group at INESC-ID Lisboa, IST / UTL.

Var	Name
x_1	DiureticsDecision
x_2	Age
x_3	Gender
x_5	SAPS-T_0
x_{10}	SOFA-T_0
x_{15}	Elixhauser Score
V_1	Propensity Score
$x_1 \cdot x_5$	-

Table 5.1: The variables used in the GP analysis.

Step 1: use an unsupervised ML technique (optional) to cluster the patients in the study group. For each cluster follow step 2.

Step 2: is divided in 2 parts:

(a) evolve a GP classifier to predict mortality as a classification problem.

(b) evolve a GP model for predicting length of stay.

5.1.1 Step 1: Unsupervised Learning of Clusters

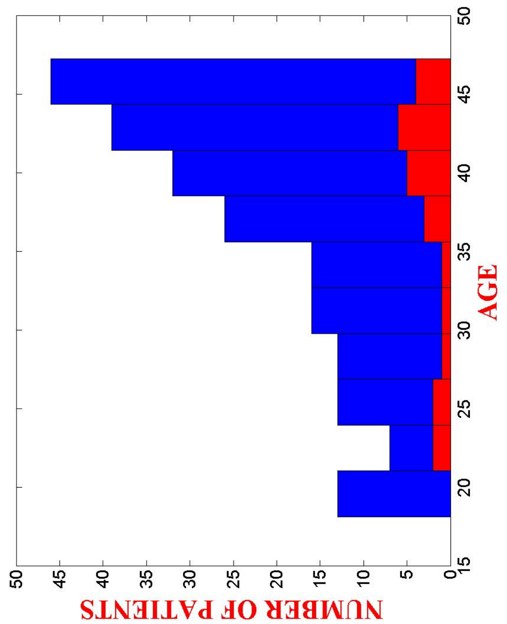
To reduce variance, clusters was (optionally) performed by using K-means[10].

Four clusters have been generated using the k-means[2] clustering method. The method was applied on a subset of the variables which describe the clinical conditions of the patients. The chosen variables were: Age, Sex, SAPS-T_0, SOFA-T_0 and Elixhauser Score.

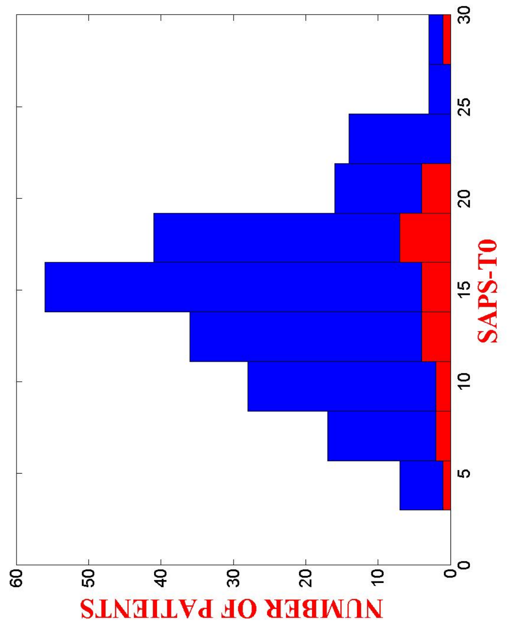
The 4 generated clusters grouped the patients according to their conditions as follows:

- **Cluster 1:** this cluster is composed by 221 patients. Age and SAPS-T_0 are shown in Figures 5.1 on the facing page.
- **Cluster 2:** this cluster is composed by 426 patients. Age and SAPS-T_0 are shown in Figures 5.2 on page 66.
- **Cluster 3:** this cluster is composed by 435 patients. Age and SAPS-T_0 are shown in Figures 5.3 on page 67.

[2] K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

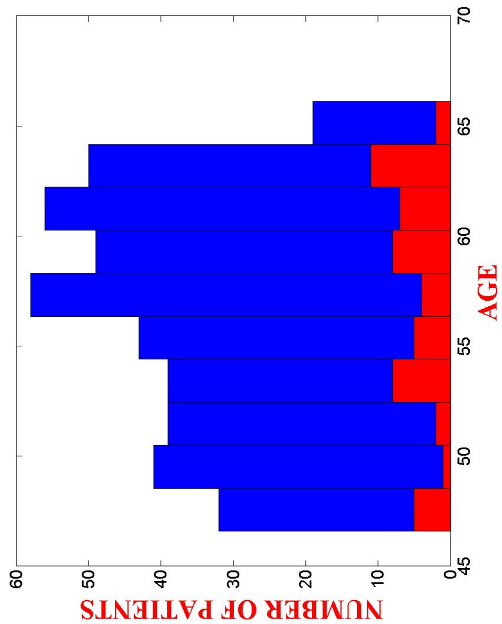


a: Histogram of Age for Cluster 1. The values are centered on 39 years old.

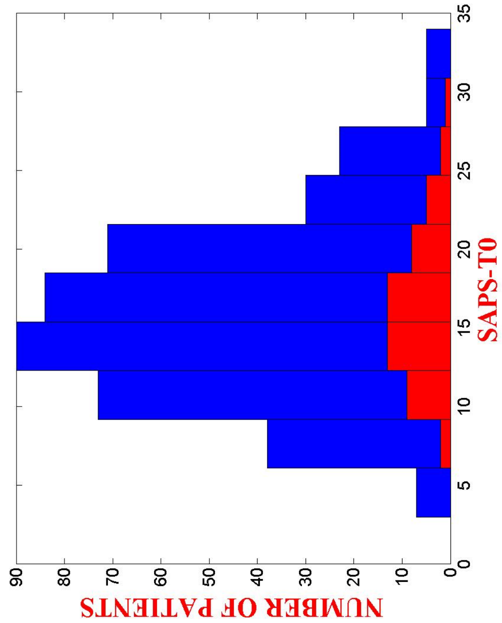


b: Histogram of SAPS-T_0 for Cluster 1. The values are centered on 15.

Figure 5.1: Histograms of Age and SAPS-T_0 for Cluster 1. In red the values for the D^+ patients only.

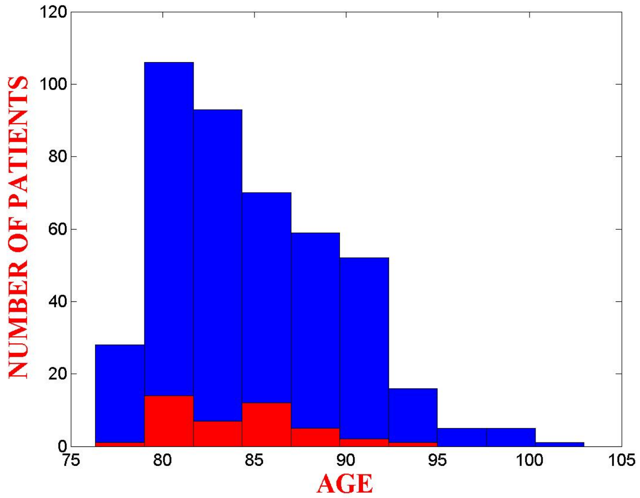


c: *Histogram of Age for Cluster 2. The values are centered on 57 years old.*

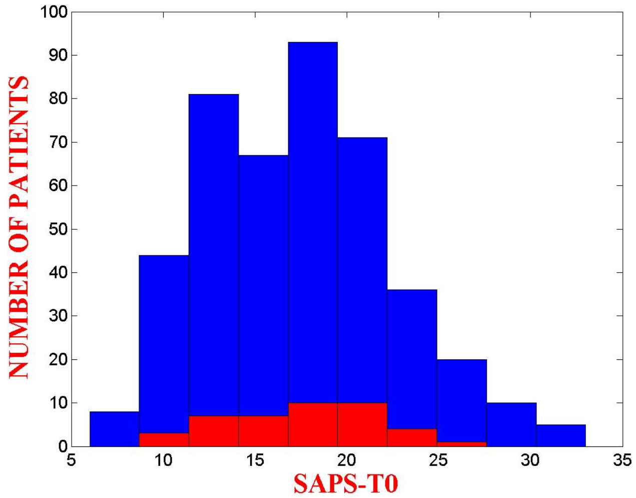


d: *Histogram of SAPS-T_0 for Cluster 2. The values are centered on 16.*

Figure 5.2: Histograms of Age and SAPS-T_0 for Cluster 2. In red the values for the D^+ patients only.



e: Histogram of Age for Cluster 3. The values are centered on 84 years old.



f: Histogram of SAPS-T_0 for Cluster 3. The values are centered on 17.

Figure 5.3: Histograms of Age and SAPS-T_0 for Cluster 3. In red the values for the D^+ patients only.

Cluster	#	D^+	D^-	Mean(SD) Age	Median Age	Mean(SD) SAPS-T_0	Median SAPS-T_0
1	221	25	196	37(8)	39	14.8(4.9)	15
2	426	53	373	56.5(5)	57	16(5.3)	16
3	435	42	393	84.8(4.6)	84	17.4(5)	17
4	440	79	371	71.5(4.2)	72	19.2(5.5)	19

Table 5.2: Clusters create by 1 execution of K-means with k=4 on variables: Age, Sex, SAPS-T_0, SOFA-T_0 and Elixhauser Score.

Operator	Value
Population Size	100
# of Generations	10
Operators	$\left\{+, -, *, /, log_2, \sqrt{}\right\}$
Probability of Reproduction p_m	0.1
Initial Probability of Crossover p_c	0.5
Initial Probability of Mutation p_m	0.5
Initialization Type	Tournament
Maximum Depth of the Trees	17

Table 5.3: Parameters of the GP executions.

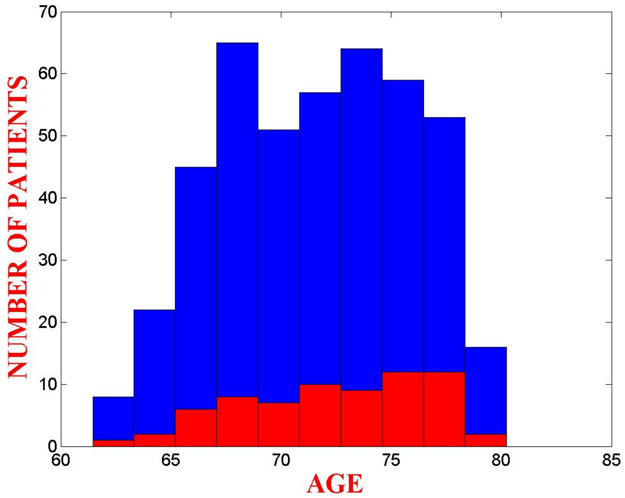
- **Cluster 4:** this cluster is composed by 440 patients. Age and SAPS-T_0 are shown in Figures 5.4 on the next page.

Clustering does not and should not create logically separated groups. However very roughly can be observed that these clusters have approximate charaterization by Age and health condition. A description of the clusters is provided in Table 5.2.

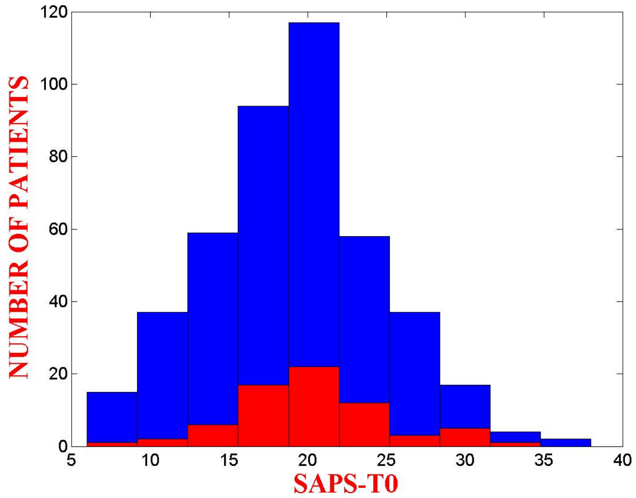
5.2 Step 2: GP modeling

For both the two outcomes, the step 2 of the analysis was performed on the whole dataset (1,522 patients), on a series of subsets composed by 4 clusters and on the less sick and sicker groups used in the analysis discussed in Chapter 4. In Table 5.3 are shown the parameters for the GP executions.

In Table 5.4 on page 70 is shown a description of the less sick and sicker groups. Median SAPS-T_0 on which the groups were splitted is 17.



g: Histogram of Age for Cluster 4. The values are centered on 72 years old.



h: Histogram of SAPS-T_0 for Cluster 4. The values are centered on 19.

Figure 5.4: Histograms of Age and SAPS-T_0 for Cluster 4. In red the values for the D^+ patients only.

Cluster	#	D^+	D^-	Mean(SD) Age	Median Age	Mean(SD) SAPS-T_0	Median SAPS-T_0
Less Sick	816	88	728	63.8(17.7)	60	13(2.9)	14
Sicker	706	101	605	68.8(15.5)	71	21.9(3.5)	21

Table 5.4: Clusters chosen according to health condition WITHOUT any machine learning technique.

Dataset	Success Rate	TP	TN	FP	FN	SEN	SPE
1522 Patients	47%	411	293	663	155	0.38	0.65

Table 5.5: GP Overall Results on the Dataset for Mortality.

5.2.1 Results on Mortality

For all the groups described above 10 indipendent GP runs have been performed using each time the 70% of the patients current group for training and the remaining 30% for testing. Each time the patients for both training and testing were randomly chosen. For all the analyzed groups the median values of the 10 runs for success rate, true positive, true negative, false positive, false negative, sensitivity ($\frac{TP}{TP+FP}$) and specificity ($\frac{TN}{TN+FN}$) will be given and discussed. All the results refer to the whole set of each group as the obtained results on the respective training and the testing sets are always similar between each others.

5.2.1.1 Results on the Original Dataset

Table 5.5 shows the overall results on the whole dataset. The models have the 46% of success on average more or less equaly divided between true positive and true negative even if the models could have problems to evaluate false positive, as the false negative are a lot. Table 5.6 shows the two best results on the whole dataset. The best model has the 63% of success.

Dataset	Success Rate	TP	TN	FP	FN	SEN	SPE
1522 Patients	63%	0	956	0	566	0	0.63
1522 Patients	59%	62	829	127	504	0.32	0.62

Table 5.6: GP Best Results on the Dataset for Mortality.

Group	Size	SR	TP	TN	FP	FN	SEN	SPE
Less Sick	816 Patients	26%	209	3	604	0	0.26	1
Sicker	706 Patients	50%	2.5	342	7	354.5	0.26	0.49

Table 5.7: GP Overall Results on the Less Sick and Sicker Groups for Mortality.

Group	Size	SR	TP	TN	FP	FN	SEN	SPE
Less Sick	816 Patients	73%	1	597	10	208	0.09	0.74
Less Sick	816 Patients	48%	116	272	335	93	0.26	0.74
Sicker	706 Patients	54%	320	64	285	37	0.53	0.63
Sicker	706 Patients	54%	320	64	285	37	0.53	0.63

Table 5.8: GP Best Results on the Less Sick and Sicker Groups for Mortality.

5.2.1.2 Results on the Less Sick and Sicker groups

Table 5.7 shows the overall results on the less sick and sicker groups. The models have the bad rate of success on average on the less sick group. In this groups the models have big problems to evaluate false positive. Table 5.8 shows the two best results on the less sick and sicker groups. The best model is for the less sick group with a success rate of 73%.

5.2.1.3 Results on the 4 Clusters

Table 5.9 shows the overall results on the 4 clusters. The models have more or less the 40% of success on average. The detection of false negative is a problem in this case too. Table 5.10 on the following page shows the two best results on the 4 clusters. The best models are the ones of clusters 1 and 2 where the success rate is better then 64%.

Cluster	Size	SR	TP	TN	FP	FN	SEN	SPE
Cluster 1	221 Patients	44%	36.5	59.5	114.5	10.5	0.24	0.85
Cluster 2	426 Patients	37%	114.5	42.5	253.5	15.5	0.31	0.73
Cluster 3	435 Patients	47%	203	0.5	230.5	1	0.47	0.33
Cluster 4	440 Patients	44%	182.5	11	244	2.5	0.43	0.81

Table 5.9: GP Overall Results on the 4 Clusters for Mortality.

Cluster	Size	SR	TP	TN	FP	FN	SEN	SPE
Cluster 1	426 Patients	71%	7	149	25	40	0.22	0.79
Cluster 1	426 Patients	66%	0	145	29	47	0	0.75
Cluster 2	440 Patients	68%	0	291	5	130	0	0.69
Cluster 2	440 Patients	65%	0	275	21	130	0	0.68
Cluster 3	221 Patients	53%	15	214	17	189	0.47	0.53
Cluster 3	221 Patients	49%	122	92	139	82	0.47	0.53
Cluster 4	435 Patients	53%	9	226	29	176	0.24	0.56
Cluster 4	435 Patients	50%	136	83	172	49	0.44	0.63

Table 5.10: GP Best Results on the 4 Clusters for Mortality.

Dataset	Median Mean Absolute Error
1522 Patients	7.5 days

Table 5.11: GP Overall Results on the Dataset for LOS.

5.2.2 Results on Length of Stay in ICU

The analysis was performed in the same way of the one for mortality, but this time the mean absolute error is the only result presented. Even in this case, all the results refers to the whole analyzed groups as the results on training and testing sets are similar between each others.

5.2.2.1 Results on the Original Dataset

Table 5.11 shows the overall results on the whole dataset. The mean absolute error is of 7.5 days on average. Table 5.12 shows the two best results on the whole dataset. The mean absolute error is more or less of 7 days on average.

5.2.2.2 Results on the Less Sick and Sicker groups

Table 5.13 on the next page shows the overall results on the less sick and sicker groups. The mean absolute error goes between 6 to 11 days on average.

Dataset	Median Mean Absolute Error
1522 Patients	6.7 days
1522 Patients	7 days

Table 5.12: GP Best Results on the Dataset for LOS.

Group	Patiens	Median Mean Absolute Error
Less Sick	816	6.4 days
Sicker	706	11.6 days

Table 5.13: GP Overall Results on the Less Sick and Sicker Groups for LOS.

Group	Patiens	Median Mean Absolute Error
Less Sick	816	5.8 days
Less Sick	816	5.8 days
Sicker	706	8.1 days
Sicker	706	8.6 days

Table 5.14: GP Best Results on the Less Sick and Sicker Groups for LOS.

Table 5.14 shows the two best results on the less sick and sicker groups. The mean absolute error goes between 5 to 8 days on average.

5.2.2.3 Results on the 4 Clusters

Table 5.15 shows the overall results on the 4 clusters. The mean absolute error goes between 5 to 10 days on average. Table 5.16 shows the two best results on the 4 clusters. The mean absolute error goes between 4 to 7 days on average.

5.2.2.4 Comment on the GP Results

Both for mortality and length of stay in ICU the prediction results are not satisfactory. Especially for length of stay the error is big. This probably is due to the difficult of the problem. The fact that the models generate on average a lot false negative indicates the difficulty of evaluating the chance of survival of the patients, expecially of the sickest ones. Furthermore, these are preliminary results and the GP models could be tuned in a better way.

Cluster	Patiens	Median Mean Absolute Error
Cluster 1	426 Patients	10.9 days
Cluster 2	440 Patients	7.9 days
Cluster 3	221 Patients	5.1 days
Cluster 4	435 Patients	8.2 days

Table 5.15: GP Overall Results on the 4 Clusters for LOS.

Cluster	Patiens	Median Mean Absolute Error
Cluster 1	426 Patients	7.3 days
Cluster 1	426 Patients	7.5 days
Cluster 2	440 Patients	6 days
Cluster 2	440 Patients	6.5 days
Cluster 3	221 Patients	4.2 days
Cluster 3	221 Patients	5 days
Cluster 4	435 Patients	6.5 days
Cluster 4	435 Patients	6.7 days

Table 5.16: GP Best Results on the 4 Clusters for LOS.

5.2.3 Simulated Outcomes

In this final analysis the chances of survival with or without diuretics have been evaluated by using the two best models produced with GP for each dataset. All the available values for the patients have been used except for the diuretics variable. Then the chances of survival for each patients have been evaluated by inserting the two possibile values, given and not given, for the diurecits variables. In this way two perfectly paired patients have been created for each actual patient.

This analysis try to overcome the real problem of an observational study that is the fact that investigators can not control the assignment of the treatments to patients and, hence, the experiment is non-randomized. With an approach of this type instead, there is the possibility of duplicate the dataset and then confront perfectly paired patients. The results of this analysis are presented in this Section. But it should be borne in mind that these results are influenced a lot by the models accuracy. So, as the models accuracy is not satisfactory, they should be considered preliminary and may be object of further analysis.

5.2.3.1 Results on Mortality

In Table 5.17 on the facing page are shows the results for mortality. An other problem that an analysis of this type could have it that the diuretics variable is binary, hence it is possible that by only flipping it, the model could not capture any difference: this could be because either the diuretics are not actually making difference or because the model is not accurate enough. In fact by analyzing the results for mortality, two things stand out: a)the models are generating a lot of false negative and the mortality rates are low,

Group	MOR 1 D-	MOR 1 D+	MOR 2 D-	MOR 2 D+
Dataset	0%	0%	0%	1%
Cluster 1	14%	14%	13%	13%
Cluster 2	11%	11%	4%	4%
Cluster 3	60%	60%	7%	7%
Cluster 4	8%	8%	70%	70%
Less Sick	13%	13%	58%	40%
Sicker	1%	0%	1%	1%

Table 5.17: GP Simulated Results for Mortality.

Group	LOS 1 D+	LOS 1 D-	LOS 2 D+	LOS 2 D-
Dataset	2 days	2 days	1.3 days	1.3 days
Cluster 1	3.6 days	3.6 days	5 days	5.6 days
Cluster 2	1.9 days	1.8 days	5.7 days	5.7 days
Cluster 3	0 days	0 days	1.5 days	1 days
Cluster 4	2.2 days	2.2 days	1.9 days	0.9 days
Less Sick	0.5 days	0.5 days	0.3 days	0.3 days
Sicker	0 days	0 days	6.6 days	6.6 days

Table 5.18: GP Simulated Results for LOS.

b)in the most of the models diuretics do not seem to make difference, for one of the reason defined above. The last thing that stands out is curiously the result for the less sick group: only in this case in fact, for the second best model, the results seems to give a better chance of survival for the patients who are getting diuretics and this goes along with the results obtained by the analysis of the first Section of this Chapter. Obviously what said is to be taken with caution, given the low accuracy of the models used.

5.2.3.2 Results on Length of Stay

In Table 5.18 are shows the results for length of stay. In this case, the results are not conclusive, possibly because of the poor accuracy of the models used.

5.2.3.3 Comments on the Results on the Simulated Outcomes

As already said, the tecnique used in this Section relies on the accurancy of the used models. In this case, the models have not a satisfactory accurancy, hence the presented results should be taken with caution. But this method could be the object of further analysis.

Chapter 6

Conclusions

6.1 Summary of Findings

A brief summary of the findings woven through Chapter 4 is now provided:

- **Finding 1: Length of stay and the administration of diuretics:**
  With respect to length of stay, health condition variables x_2, x_5 and x_{15} (Age, SAPS-T_0, and Elixhauser Score) have statistically significant effects, as does propensity score. The null hypothesis that the independent effect of diureticsDecision is not significant on length of stay, is rejected (p-value < 0.001).
  For length of stay outcome, these findings imply health condition is not a confounding factor and diureticsDecision is independently significant in its effect on length of stay.
  This validates the findings of the quintile analysis for length of stay in Chapter 3. They indicate a statistically significant difference in length of stay for quintile 1 ($PS \in [0.00; 0.01]$), quintile 3 ($PS \in [0.04; 0.08]$) and quintile 5 ($PS \in [0.19; 0.99]$), leading to the conclusion, qualified for this study group, that the administration of diuretics increases a patient's length of stay in the ICU.

- **Finding 2: Independent Effect of the administration of diuretics on mortality:**
  The null hypothesis that the independent effect of diureticsDecision is not significant on mortality, is accepted (p-value > 0.05). Hence, the diureticsDecision does not have a statistically significant independent effect on mortality.

- **Finding 3: Cross-Dependent Effect of the administration of diuretics and health condition on mortality:**

The null hypothesis that SAPS-$t0$_diureticsDecision has a not significant cross-dependent effect on mortality is rejected given the p-value = 0.013. Through adjusted regression analysis with ModelC.Sicker and ModelC.Sicker, (see Table 4.2 on page 56 columns 4-5 and 6-7), the null hypothesis that diureticsDecision has a not significant cross-dependent effect on mortality in the less sick subset is rejected. The null hypothesis that diureticsDecision has a not significant cross-dependent effect on mortality in the sicker subset is *NOT* rejected.
Furthermore, through health condition adjusted stratification analysis, see Table 4.3 on page 60, in the less sick subset of quintile 4 ($PS \in [0.06; 0.15]$) mortality rate is significantly less for the patients with the administration of diuretics compared to those without. It is not significantly less for quintile 5 ($PS \in [0.15; 0.99]$). Per Table 4.4 on page 61 mortality rate is not significantly different for either of quintiles 4 or 5.

In Chapter 5 a preliminary analysis using genetic programming is described. The Chapter's contribution is to outline the method, whereas the produced results are not reliable.

6.2 Future Work

6.2.1 Propensity Analysis

The primary objective of this work was to develop a statistical methodology based on propensity analysis and logistic or linear regression. A study group was selected for the analysis and results on it were produced. A first set of possible future work would involve reformulating the study group both with the aim of studying a larger better selection of patients who exhibit high fluid levels (than has been done in this work) but a different pathology.

It has to be said that during the definition of the study group and during the subsequent extraction from the Mimic II Clinical Database, certain choices were made, as described in the respective Chapters. This was necessary given the vastness of the topic. Hence, it could be of interest to further study possible alternatives to the already explored choices. While the study of other diseases with the same method would be of obvious interest: for this purpose all the developed procedures were designed to be easily reused in this context.

6.2.2 GP Analysis

As regards the analysis carried out by genetic programming, its aim was to be only a first exploration of the method. This entire analysis would lend itself to further study. First of all, it requires a precise study of the configuration used for the executions of the algorithm with the aim to improve the reliability of the produced models.

If this is done, the updated results could be re-evaluated and potentially prompt new experiments or method refinements. For example, generating different models using separately the group of patients to which diuretic was administered and the one who did not get diuretics and even further dividing the groups by illness in the case of mortality.

Furthermore, different machine learning techniques for both the generation of the clusters and of the models should be tried and compared.

Bibliography

[1] David M. Mannino M.D. Stephanie Eaton M.D. Greg S. Martin, M.D. and M.D. Marc Moss. The epidemiology of sepsis in the united states from 1979 through 2000. *The New England journal of medicine*, 2003.

[2] Stanley Lemeshow PhD Jean-Roger Le Gall, M.D. and M.D. Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Journal of American Medical Association, 270:29572963, 1993*, 1993.

[3] Takala J Willatts S De Mendona A Bruining H Reinhart CK Suter PM Thijs LG. Vincent JL, Moreno R. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Med (1996) 22:707-710*, 1996.

[4] FRCPC MSc Peter C. Austin PhD Alison Jennings BSc MSc Hude Quan MD PhD Carl van Walraven, MD and FRCPC MSc Alan J. Forster, MD. A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. *Medical Care, Volume 47, Number 6*, 2009.

[5] Mauricio Villarroel Gari D. Clifford1, Daniel J. Scott. User guide and documentation for the mimic 2 database, 2012.

[6] PAUL R. ROSENBAUM and DONALD B. RUBIN. Reducing bias in observational studies using subolassification on the propensity score. *Journal of the American Statistical Association*, 1984.

[7] Peter C. Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine 2008; 27:20372049*, 2008.

[8] GreenFacts. http://www.greenfacts.org/glossary/abc/confounding-factor.htm.

[9] Sara Silva. Gplab - a genetic programming toolbox for matlab, April 2007.

[10] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *University of California Press*, 1967.

[11] F B Cerra R P Dellinger A M Fein W A Knaus R M Schein R C Bone, R A Balk and W J Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest 1992;101;1644-1655*, 1992.

[12] FCCP; Mitchell P. Fink MD FCCP; John C. Marshall MD; Edward Abraham MD; Derek Angus MD MPH FCCP; Deborah Cook MD FCCP; Jonathan Cohen MD; Steven M. Opal MD; Jean-Louis Vincent MD FCCP PhD; Graham Ramsay MD; Mitchell M. Levy, MD. 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Crit Care Med 2003 Vol. 31, No. 4*, 2001.

[13] W. P. Weber M. Adamina, U. Guller and D. Oertli. Propensity scores and the surgeon. *British Journal of Surgery 2006; 93: 389394*, 2006.

[14] Unitá Operativa di Epidemiologia e Biostatistica. http://www.ospedalebambinogesu.it/Portale2008/Default.aspx?Iditem=1178.

[15] Gregory Piatetsky-Shapiro William J. Frawley and Christopher J. Matheus. Knowledge discovery in databases: An overview. *AI Magazine (Vol 13, No 3), 57-70*, 1992.

[16] Gregory Piatetsky-Shapiro Usama Fayyad and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine (Vol 17, No 3), 37-54*, 1996.

[17] Rodolph A. Miller. Medical diagnostic decision support systems - past, present, and future. *Journal of the American Medical Association*, 1994.

[18] Heather McDonald M. Patricia Rosas-Arellano P. J. Devereaux Joseph Beyene Justina Sam R. Brian Haynes Amit X. Garg, Neill K. J. Adhikari. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. *American Medical Association*, 2005.

[19] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[20] T. Mitchell. The discipline of machine learning, 2006.

[21] Leonardo Vanneschi. *Theory and Practice for Efficient Genetic Programming*. PhD thesis, University of Lausanne, 2004.

[22] David B. Fogel. What is evolutionary computation? *IEEE Spectrum*, 2000.

[23] John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. MIT Press, 1992.

[24] D E Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[25] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. 1992.

[26] John R. Koza. *Genetic programming II - Automatic Discovery of Reusable Programs*. 1994.

[27] David Andre Martin A. Keane John R. Koza, Forrest H. Bennett III. Genetic programming iii - darwinian invention and problem solving. *IEEE Transactions on Evolutionary Computation, Vol. 3, NO. 3, September 1999*, 1999.

Appendix A

Medical Backgrounds

This Appendix aims at providing an overview on medical concepts usefull to understand the analysis discussions. In particular will be defined what sepsis is. The Appendix does not aim at being a full medical guide on the topic, but intends to provide some useful basic medical knowledge.

A.1 Definition of Septis

Sepsis is the leading cause of death in noncoronary intensive care units (ICU) in the United States[11].

It is a potentially deadly medical condition that is characterized by a whole-body inflammatory state, called a systemic inflammatory response syndrome or SIRS, and the presence of a known or suspected infection. The body may develop this inflammatory response by the immune system to microbes in the blood, urine, lungs, skin, or other tissues. Severe sepsis is the systemic inflammatory response, plus infection, plus the presence of organ dysfunction[1].

The core of the current definition of sepsis arose from the 1991 American College of Chest Physicians / Society of Critical Care Medicine (ACCP / SCCM) Consensus Conference. This definition was revisited and slightly modified by the 2001 Internal Sepsis Definition Conference.

A.1.1 1991 ACCP / SCCM Consensus Conference

An American College of Chest Physicians / Society of Critical Care Medicine Consensus Conference was held in Chicago in August 1991 with the goal of

[1]Organ dysfunction is a condition where an organ does not perform its expected function. When the organ dysfunction gets bad to such a degree that normal homeostasis cannot be maintained without external clinical intervention, occurs organ failure.

Abnormalities	**Values**
Temperature	<36 °C (96.8 °F) or >38 °C (100.4 °F)
Heart rate	>90/mins
Respiratory rate	>20/min or $PaCO_2$<32 mmHg (4.3 kPa)
WBC	<4x10^9/L (<4000/mm^3), >12x10^9/L (>12,000/mm^3) or 10% bands

Table A.1: According to the 1991 ACCP / SCCM definition, SIRS is diagnosed when a patient has two or more of the clinical abnormalities.

agreeing on a set of definitions that could be applied to patients with sepsis and its sequelae[11]. The conference provided a set of definitions used to characterize the progression of the disorder.
Sepsis refers to a clinical spectrum of complications starting with the initial infection and ultimately progressing to septic shock. It initially manifests as the nonspecific systemic inflammatory response syndrome (SIRS). SIRS is diagnosed when a patient has two or more of the clinical abnormalities provided in Table A.1. The patient must present at least two of the following SIRS abnormalities: temperature, heart rate, respiratory rate, WBC[2].
As it has been said, according to the American College of Chest Physicians / Society of Critical Care Medicine, there are different levels of sepsis:

- Sepsis: defined when SIRS occurs and there is a documented or highly suspected infection.
- Severe sepsis: defined as sepsis with organ dysfunction, hypoperfusion[3], or hypotension[4].
- Septic shock: defined as sepsis with refractory arterial hypotension or hypoperfusion abnormalities in spite of adequate fluid resuscitation.

The progression of sepsis symptoms is shown in Figure A.1 on page iii.

[2]Total white blood cell count.
[3]Decreased blood flow through an organ.
[4]Abnormally low blood pressure, especially in the arteries of the systemic circulation.

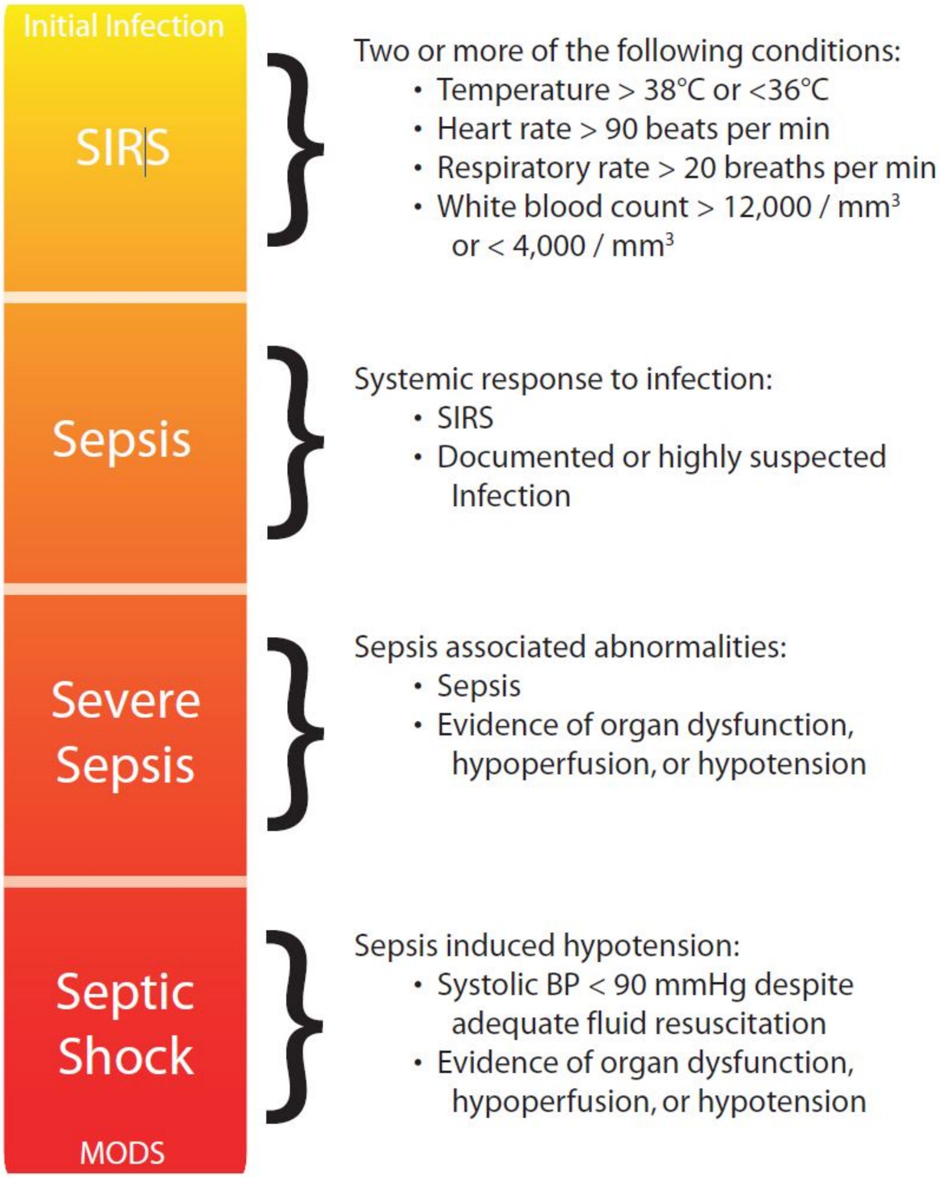


Figure A.1: The clinical spectrum of sepsis begins with the nonspecific systemic inflammatory response syndrome and progresses through increasing inflammatory response stages. The spectrum ultimately ends in septic shock and/or multiple organ dysfunction syndrome (MODS).

A.1.2 2001 Internal Sepsis Definition Conference

Ten years after the 1991 ACCP / SCCM Consensus Conference was held to establish to uniform definitions for sepsis and the associated spectrum of progressive injurious processes, the 2001 Internal Sepsis Definition Conference revisited these definitions to evaluate their efficacy and suggest improvements. In the conference was stated that there had been an impetus from experts in the field to modify these definitions to reflect the current understanding of the pathophysiology of these syndromes[12].
Participants of the 2001 Internal Sepsis Definition Conference agreed that in the 1991 ACCP / SCCM Consensus Conference, SIRS definition was overly sensitive and provided little clinical utility in the initial diagnosis of sepsis. Clinicians did not make the diagnosis of sepsis based on the 1991 SIRS criteria, but rather by analyzing the host of symptoms and deciding the patient looks septicregardless of a documented source of infection[12].
Thus in hopes to increase utility in making the sepsis diagnosis, a more comprehensive list of SIRS criteria was established as provided in Table A.2 on page v. Except for expanding the SIRS list, the conference found no evidence to support any need for changes in the 1991 ACCP / SCCM Consensus Conference definition.

A.2 Epidemiology

In the United States, sepsis is the second-leading cause of death in non-coronary Intensive Care Unit (ICU) patients and the tenth most common cause of death overall according to data from the Centers for Disease Control and Prevention (the first being heart disease)[1]. Sepsis is common and also more dangerous in elderly, immunocompromised, and critically ill patients. It occurs in 12% of all hospitalizations and accounts for as much as 25% of ICU bed utilization. It is a major cause of death in intensive-care units worldwide, with mortality rates that range from 20% for sepsis, through 40% for severe sepsis, to over 60% for septic shock.
It is importante to note that the results in the studies of sepsis are highly sensitive to the case definition for sepsis used in the study. Additionally, retrospective studies (for examples using discharge summaries) are at the mercy of clinicians to make diagnoses and most of those are made on the basis of a gut feeling that the patient is looking septic.

Diagnostic criteria for sepsis **Infection, documented or suspected, and some of the following**
General variables: Fever (core temperature >38.3 °C) Hypothermia (core temperature <36 °C) Heart rate >90min^{-1} or >2 *SD* above the normal value for age Tachypnea[5] Altered mental status Significant edema or positive fluid balance (>20 mL/kg over 24 hrs) Hyperglycemia (plasma glucose >120 mg/dL or 7.7 mmol/L) in the absence of diabetes
Inflammatory variables: Leukocytosis (WBC count >12,000 μL^{-1}) Leukopenia (WBC count <4000 μL^{-1}) Normal WBC count with >10% immature forms Plasma C-reactive protein >2 *SD* above the normal value Plasma procalcitonin >2 *SD* above the normal value
Hemodynamic variables: Arterial hypotension (SBP <90 mm Hg, MAP <70, or an SBP decrease >40 mm Hg in adults or <2 *SD* below normal for age) S$\bar{v}O$2 >70%b Cardiac index >3.5 L$\cdot$ min$^{-1}\cdot$ M^{-23}
Organ dysfunction variables: Arterial hypoxemia (PaO_2/FIO_2<300) Acute oliguria (urine output <0.5 mL$\cdot$ kg$^{-1}\cdot$ hr^{-1} or 45 mmol/L for at least 2 hrs) Creatinine increase >0.5 mg/dL Coagulation abnormalities (INR >1.5 or aPTT >60 secs) Ileus (absent bowel sounds) Thrombocytopenia (platelet count <100,000 μL^{-1}) Hyperbilirubinemia (plasma total bilirubin >4 mg/dL or 70 mmol/L)
Tissue perfusion variables: Hyperlactatemia (>1 mmol/L) Decreased capillary refill or mottling

Table A.2: In the 2001 Internal Sepsis Definition Conference the definition of SIRS was updated.

Appendix B

Software

All the procedures used in this work will be described in this Appendix. The first Section refers to the extraction from the Mimic II Clinical Database of the variables for the patients in the study group and expands the discussion made in Chapter 2. The second Section describes the variables preparation modules and it is also an extension of the discription provided in Chapter 2.

The third Section describes all the procedures used to perform the propensity analysis described in Chaper 3.

The last Section briefly summarize the procedures regarding the outcome analysis and the machine learning with gp analysis of Chapters 4 and 5.

B.1 Dataset Extraction

As already anticipated in Chapter 2, the extraction of the records for the analysis has been performed by three Matlab Scripts: a)SQL Script, b)Diuretics Naive Condition and c)Data Filtering. In this Section a deeper description of the produced code will be made.

B.1.1 SQL Script

The SQL Script module consists in 21 queries on a PostgreSQL database containing an updated image of the Mimic II Clinical Database. The schema of the database is well defined in[5], so refers to it for a deeper description. The next contents are also drawn from[5].

In Figure B.1 on page viii are shown the relationships between the tables of the database which identify a patient. The clinical conditions and related exams of each patient are stored for four significant contexts each of these in a separated series of tables: chart events, see B.2 on page ix, medication

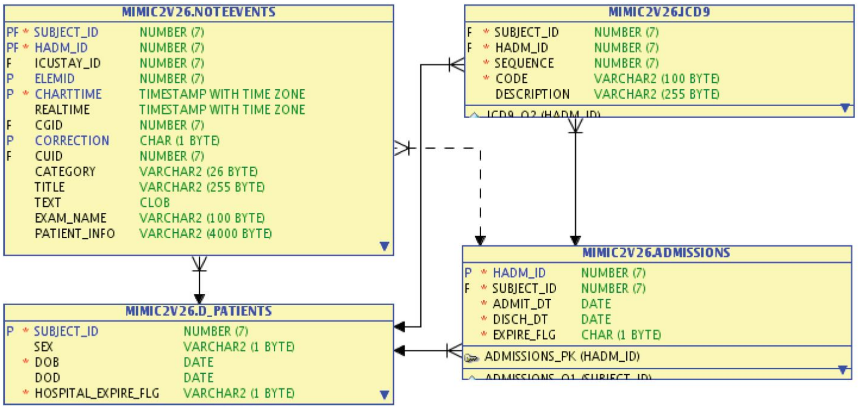


Figure B.1: Relationship between the table containing the patients' data and hospital admissions, ICD9 codes and note events tables.

events, see B.3 on page x, input/output events, see B.4 on page xi and lab events, see B.5 on page xii.

Should be noted that even if different tables are made for each context, all of them have a central table where the timeline of the events for each patients are saved and a series of other tables where are saved the descriptions of these events, for instance the kind of performed medication or the duration of an exam.

The queries are performed for the following tasks and realize the extraction of the values for the patients on the cohort of study:

1. **triples ordered by subject_id:** performs the steps of the filtering and order the results by subject_id;

2. **triples ordered by hadm_id:** performs the steps of the filtering and order the results by hadm_id;

3. **triples ordered by icustay_id:** performs the steps of the filtering and order the results by icustay_id;

4. **discharges summaries ordered by hadm_id:** extracts the discharge summaries and order the results by hadm_id;

5. **diuretics ordered by icustay_id:** extracts the IDs of the patients who got diuretics at least one time during the stay in ICU and order the results by icustay_id;

6. **diuretics first time ordered by icustay_id:** extracts the time when the patients got diuretics for the first time and order the results by icustay_id;

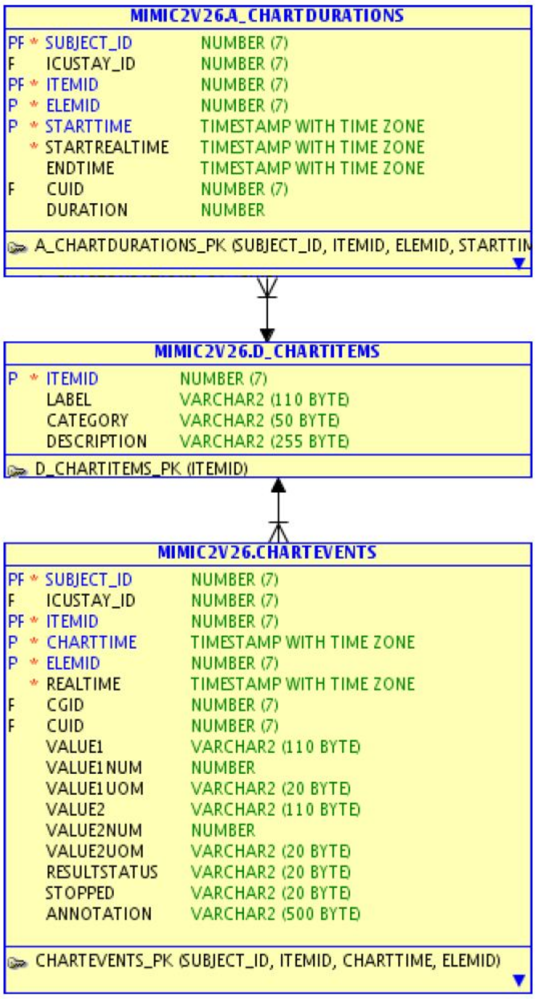


Figure B.2: Patients' chart values are stored in 3 tables: chartevents, d_chartitems and a_chartdurations. The events' timeline is in the chartevents table while the related elements and durations are in the other tables

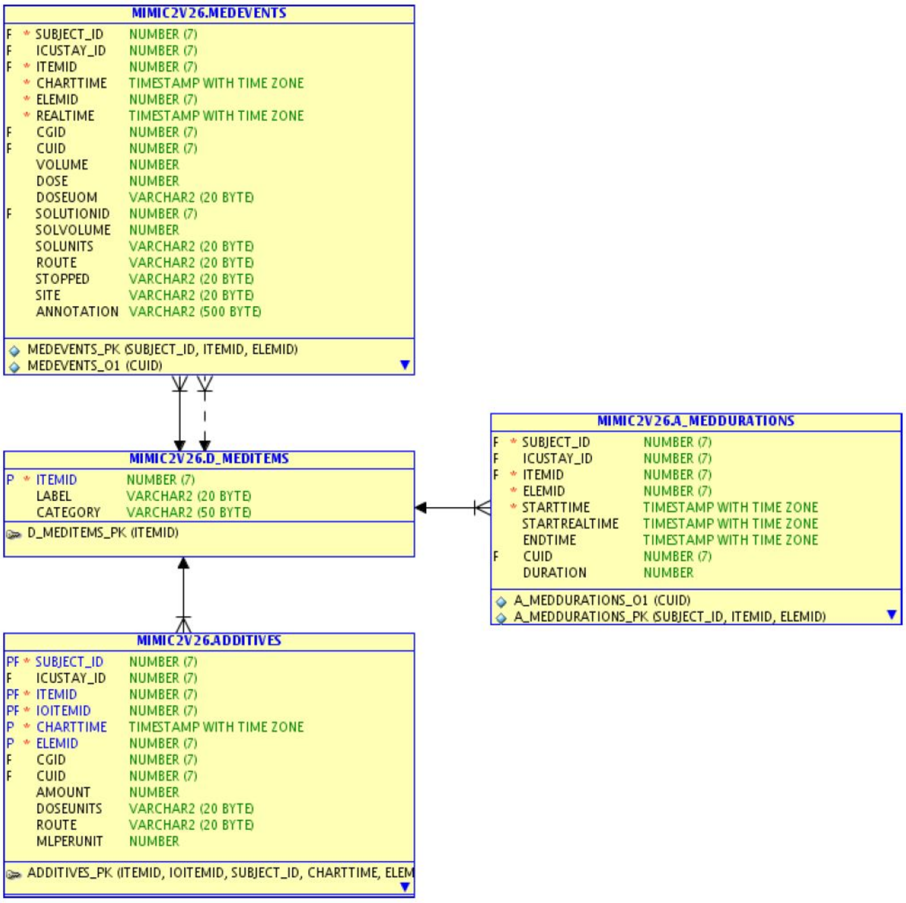


Figure B.3: Patients' medications are stored in 4 tables: medevents, d_meditems, a_meddurations and additives. The events' timeline is in the medevents table while the related elements and durations are in the other tables.

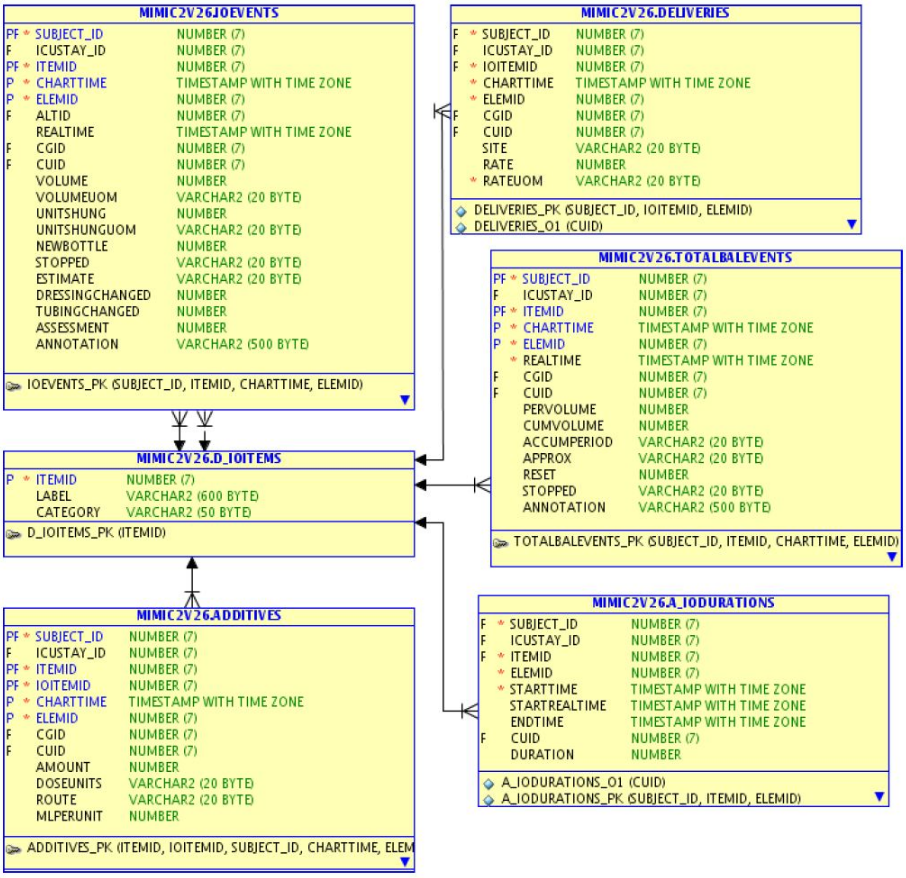


Figure B.4: Patients' IO values are stored in 6 tables: ioevents, d_ioitems, a_iodurations, deliveries, totalbalevents and additives. The events' timeline is in the ioevents table while the related elements and durations are in the other tables.

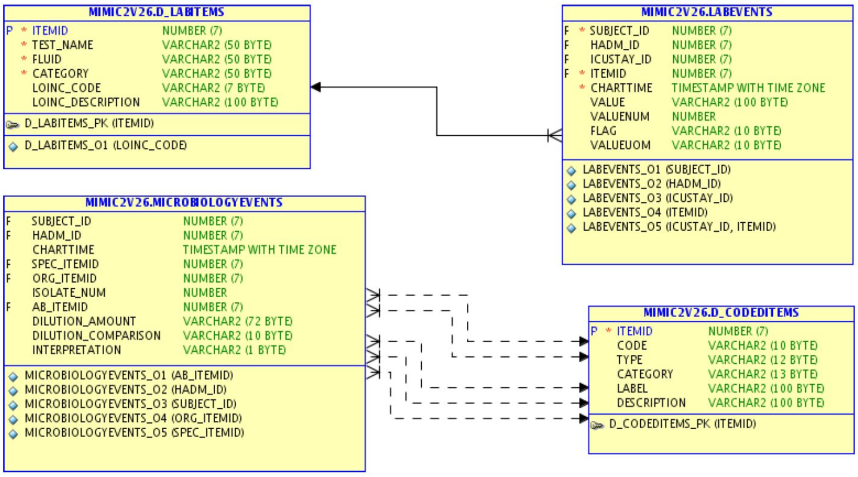


Figure B.5: *Laboratory and microbiology tests are stored in 4 tables: labevents, microbiologyevents, d_labitems and d_coded items. The events' timeline is in the labevents and microbiologyevents tables while the related elements, containing full descriptions of the lab tests (with LOINC codes, a database and universal standard for identifying medical laboratory observations) and microbiology tests (specimin, organism and antibiotic), are in the other tables.*

7. **demographic data ordered by icustay_id:** extracts age and gender for each patient and order the results by icustay_id;

8. **race ordered by hadm_id:** extracts the race for each patient and order the results by hadm_id;

9. **saps ordered by icustay_id:** extracts the saps score for each patient and order the results by icustay_id;

10. **sofa ordered by icustay_id:** extracts the sofa score for each patient and order the results by icustay_id;

11. **elixhauser ordered by hadm_id:** extracts the elixhauser score for each patient and order the results by hadm_id;

12. **elixhauser binary ordered by hadm_id:** extracts the elixhauser parameters for each patient and order the results by hadm_id;

13. **creatinine ordered by icustay_id:** extracts the creatinine values for each patient and order the results by icustay_id;

14. **fluids inputs ordered by icustay_id:** extracts the fluids inputs values for each patient and order the results by icustay_id;

15. **fluids outputs ordered by icustay_id:** extracts the fluids outputs values for each patient and order the results by icustay_id;

16. **use of vasopressors ordered by icustay_id:** extract a binary value representing if vasopressors were given to the patient during the ICU. For the positive records are save the IDs of the patients order by icustay_id;

17. **mechanical ventilation ordered by icustay_id:** extract a binary value representing if patient was on mechanical ventilation during the ICU. For the positive records are save the IDs of the patients order by icustay_id;

18. **arterial bp ordered by icustay_id:** extract the bloop pressure values and order the results by icustay_id;

19. **arterial bp mean ordered by icustay_id:** extract the bloop pressure mean values and order the results by icustay_id;

20. **mortality within 30 days ordered by icustay_id:** extract the mortality value and order the results by icustay_id;

21. **length of stay ordered by icustay_id:** extract the lengths of stay in ICU and order the results by icustay_id;

The first three queries perform the filtering steps and order the results by the three IDs. This have been done to reduce the complexity of the next procedure, that will have to perform some kind of searching in a list of sorted files. The three files generated by these queries save the three IDs of the records of the dataset and in this sense represent the individual of a possibile GP population applied to the problem. The query four extract the discharge summaries used in the next procedure.

The following queries extracts the data for each variable used in the further the analysis. For the variables which have a timeline (see queries 6, 9, 10, 13, 14, 15, 18, 19), the time when a single value is refferring to is saved has an offset with respect to when the respective patient entered the ICU. For those variables with timeline the sampling rate of when the values are saved is irregular, to normalize the rate to a daily one have been computed the average of the values considered by day. Furthermore, it happens that sometimes there are more values at the same time. To decrease the weight of outlayes, have been computated the median value of those values.

The queries five and six extract the needed values for the input, that is diuretics given or not in the ICU, variable. The time of the first dose of

diuretics is extracted for statistical analysis on when this first dose occurs. Queries seven and eigh extract pieces of information about age, gender and race.

Queries nine, ten and eleven extract the values for saps, sofa and elixhauser, the twelfth query extract a binary value for 9 of the 30 parameters of the elixhauser score. Query thirteen extract the values for the creatinine variable.

Queries fourteen and fifteen extracts the amounts of fluids inputs and outputs administrated to each patients. As these values are amount, the single values available are summed daily instead of computing the average.

Queries sixteen and seventeen extract two binary values. The first one is the use of vasopressors during the stay in ICU and the second one capture a binary value regarding the usage of mechanical ventilation for a patient during the ICU stay. This second value is not directly available in the Mimic II Clinical Database, therefore to obtain this value has been used an euristic procedure: if there are two changes in the ventilator's parameter for a patient at a distance longer than 6 hours, have been assumed that the current patient went from extubated to intubated, hence the mechanical ventilation was considered occurred. Queries eighteen and nineteen extract the values for the bood pressure.

The last two queries, the twenty and twenty first, extract the values for the two identified outcomes: mortality within 30 days and the length of stay in the ICU.

B.1.2 Diuretics Naive Condition

The diuretics naive condition refers to the fact that a certain patients received diuretics before the admission in the ICU. A patients is considered naive if didn't receive any kind of diuretics before entering the ICU, all the patients that were not naive, were discarded from the dataset.

The condition was verified by parsing a text like field extracted from the Mimic II Clinical Database, the discharge summary. Everytime a patient leave the hospital, a summary is stored which a series of information regarding the stay of the patient in the hospital, the drugs the patients declared to have received before entering the hospital and the medications given to the patient while leaving.

However, the discharge summaries don't have a standard form: the Section of the summary are usualy, even if not always, demarcated by Section titles. These titles, though, are not always the same. For instance the diuretics informstion regarding the drugs given to the patient before entering

the ICU could have been demarcated with *DRUGS ON ADMISSION* or with *ON ADMISSION* or in other different ways.

The parsing process was made with a Perl script that gets as an imput a file with the HADM_ID and the summary of each patient and provides as an output a file with the list of HADM_ID of the naive patients.

B.1.3 Data Filtering

Aim of this procedure is to create a list of records combining the SQL filtering to the diurecits naive condition. Plus the procedure set a series of variables to be mandatory for a patient, for instance the fluids inputs or outputs, and discard the records without a value for them.

The objective is achived in two steps. The first one go through the files considered mandatory and save three files containing the triples ordered by one of the three IDs each. The diuretics naive condition is combined with the mandatory variables. The second step go throught the files provided by step one and save in series of files the data for all the variales.

The input of this procedure are the files provided by the previous ones except for the discharge summary, which are analized in the diretics navice condition procedure. The output is a series of files all of them ordered by the respective ID, three with the triples (that is defining the dataset's records) and the others with all the variables.

An easy method to go through the files for the seach would be to use two nested loops for each file, one going through the triples file ordered by the ID of the current variable, and one going through the file of the variable: this approach has a complexity $O[N \cdot (n \cdot m)]$, being N the number of variables to be analized, n dimention of input file that is the current variable file, and m the dimension of the dataset. In this way the extraction process was too computationally expensive.

The complexity was then reduce to $O[N \cdot (n + m)]$ by exploiting the fact that the files were in numerical acending order. In this way every line of the files were read to most one time each. The pseudo-code of the algorithm is shown in Algorithm 1 on page xvi.

B.2 Variables Preparation

The variables preparation consists in a transformation of the dataset's values in a format straight forward usable in the analysis procedures.

Before the variables preparation itself, are executed a few script as follow:

Algorithm 1 Pseudocode of the algorithm used to perform the merging process.

external loop go through the IDs file ($F1$)
internal loop go through the current variable file ($F2$)
while (Both files have lines) **do**
  if $F1 == F2$ **then**
    go on reading one line from $F2$ file. Then write one line to output
  end if
  if $F1! = F2ANDF1 < F2$ **then**
    save $F2$ data and go on one step with $F1$
  end if
  if $F1! = F2ANDF1 > F2$ **then**
    go on reading from file $F2$
  end if
end while

- run a script to get age and gender values: in the Mimic II Clinical Database all patients over 90 years old have been saved as 200 years old, the script get the true age. Then the value for gender, that is M for male and F for female, is modified in a binary one, 0 for male and 1 for female;

- run a script to get the elixhauser values: extracts in single files the binary values for the exlixhauser parameters;

- run a script to compute the balance values defined as $inputs - outputs$;

- run a script to define the times where to save certain values, that is $T1$, $T2$ and $T3$, the times discussed previously in Chapter 2.

Completed these steps, the variables preparation starts and each variable is saved in a file, in particulare the following procedures have been realized:

- average processing: to save the values for the variables that require to calculate the average (or the sums) till a certain day;

- binary processing: to save the values for the binary variables;

- numeric processing: to save the values for the numeric variables;

- time processing: to save the values for the variables for the variables with timelines.

The scripts define above, also perform an analysis on the results and save the in a file.

After the variables preparation the procedure save the results in two files directly usable for the analysis. The first file contains the labels for the computated variables, the second the data itself.

B.3 Propensity Analysis

The propensity analysis, following[6], have been performed by the following procedures:

1. **Fitting the Propensity Score:** the propensity score, that is the the conditional probability of assignment to a particular treatment given a vector of observed covariates, is calculated with a logit model.

2. **Generating the five Quintile:** the patients are ranked according their propensity score and then divided in five quintiles.

3. **Assessing the Balance:** the balance in each of the five quintilesis evaluated with the ANOVA test for primary effects and secondary effects.

4. **Refining the Quintile:** the balance in the quintileis improved by inserting the variables with large F-ratio and the computing a new logit model with them.

In Figure B.6 on page xviii the basic elements of the code produced to implement the propensity method.

Now a description of these procedures will be provided. These Sections extend what have already been discussed in Chapter 3. In Algorithm 2 on page xix is shown an overview of all the propensity score process.

B.3.1 Fitting the Propensity Score

The propensity score is calculated by performing a two-phases stepwise discriminant analysis with a logit model: first it is generated a model evaluating the main effects of all the variables in the dataset. The output of this process is a list of variables chosen as main effects. In the study 11 variables were chosen in this first logit model.

After that, a second stepwise discriminant analysis is performed considering only the variables whose main effects were chosen in the first logit model and in this second analysis the interactions between these variables

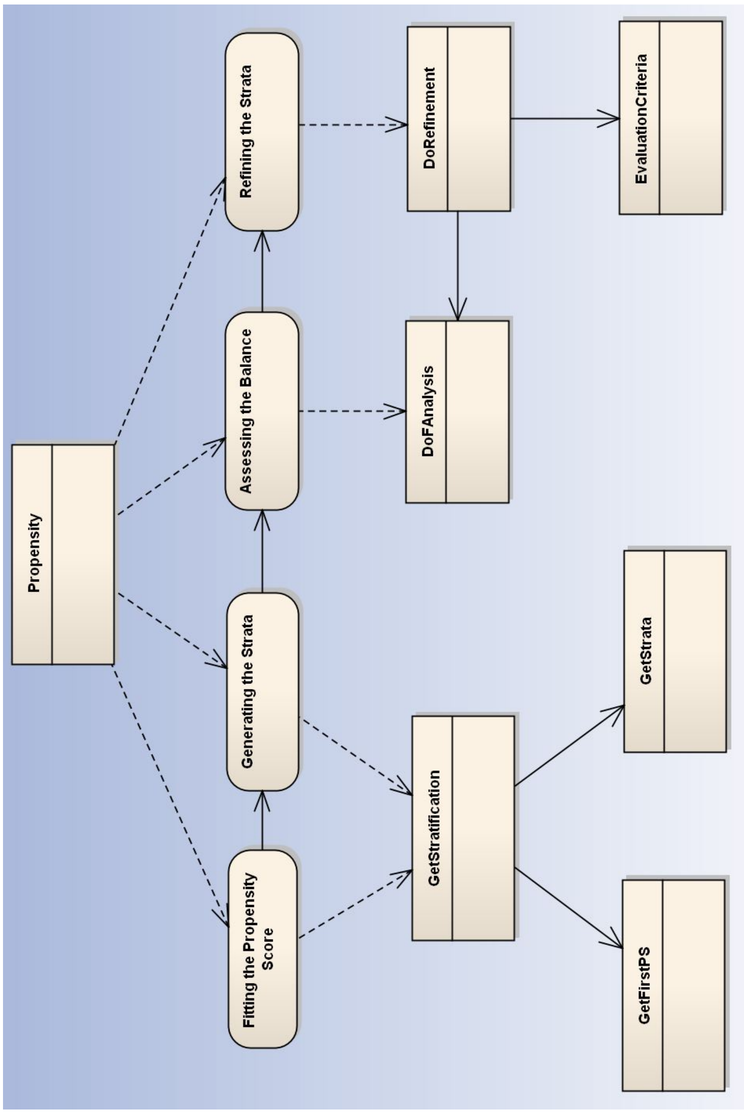


Figure B.6: Steps of the propensity analysis. It is composed by four importants parts: a)fitting the propensity score, b)generating the five quintile, c)assessing the balance and d)refining the quintile.

Algorithm 2 Pseudocode of the Propensity Score process.

Generate the stepwise logit model $M1$ on main effects
Generate the stepwise logit model $M2$ including interactions
Based on $M2$, compute the propensity score PS
Rank the patients based on PS and perform the subclassification
Evaluate the balance according to the F-ratios
Rank the variables based on their F-ratios
Chose the first variable in the ranked list, X
while (All the variables not in $M2$ have not considered) **do**
  if (Including X), its F-ratio improves **then**
    include X in the model and proceed to the next variable
  end if
  if (Including X^2), its F-ratio improves **then**
    include X^2 in the model and proceed to the next variable
  end if
  if (Including interaction of X), its F-ratio improves **then**
    include interaction of X in the model and proceed to the next variable
  end if
  Chose the next variable in the ranked list, $X = X_i$
end while

are considered: the result of this second logit model was a model of 17 variables, 11 main effects defined in the first stepwise discriminant analysis and 6 interactions between them added in the second stepwise discriminant analysis. The full list of those variables is presented in Chapter 3.

At this point, using this model it is possibile to calculate a probability of getting the diuretics. Being x the resulting values available for each patient using the model on their data, $p(x)$ of receiving diuretics is: $p(x) = \frac{e^x}{1+e^x}$.

B.3.2 Generating the five Quintile

The records of the patients are then ranked in an increasing order according to the propensity score calculated with the second logit model. Based on this ranked order, are define 5 groups, called quintile, and the patients with a lower probability of being administrated with diuretics are in group 1 while the ones with the higher one are in group 5.

B.3.3 Assessing the Balance

The balance in each quintileis evaluated using the ANOVA test. ANOVA provides a statistical test of whether or not the means of several groups are all equal.

The test produces a ration $F = \frac{variance-between-groups}{variance-within-groups}$, it is based on the partitioning of the total sum of squares S into components related to the effects used in the model.

Will be defined as S_1 the sum of squares of the differences between the means in each group, m_i, and the overall mean m, that is: $S_1 = \sum_i n_i \cdot (m_i - m)^2$, with n_i number of elements of group i.

The will be defined as S_2 the sum of squares of the differences between the means in each group, m_i, and the value of a certain element of that group, $x_{i,j}$. That is $S_2 = \sum_i \sum_j n_i \cdot (x_{i,j} - m_i)^2$.

At this point, being k the number of groups to be evaluated and n all the elements in the groups, the F-ratio defined by ANOVA is $F = \frac{(\frac{S_1}{k-1})}{(\frac{S_2}{n-k})}$.

B.3.3.1 Evaluating the Balance

At this point, the one-way ANOVA test have been performed on the original dataset: the test compares the treated vs not treated patients and, of course, the F-values are high.

Then the two-ways ANOVA is calculated on the dataset divided in the 5 quintile. The test produces two values: the first one is made considering

the main effect of the diuretics (given vs not given) variable. Consider for instance the comparison between two binary variables, if the combinations of their values are listed in a table, the result is a 2×2 matrix: the main effect values that the two-way ANOVA would calculate on this table are two, the first one considering the rows and the second one considering the columns. The main effect values in the ANOVA test consider as gourps

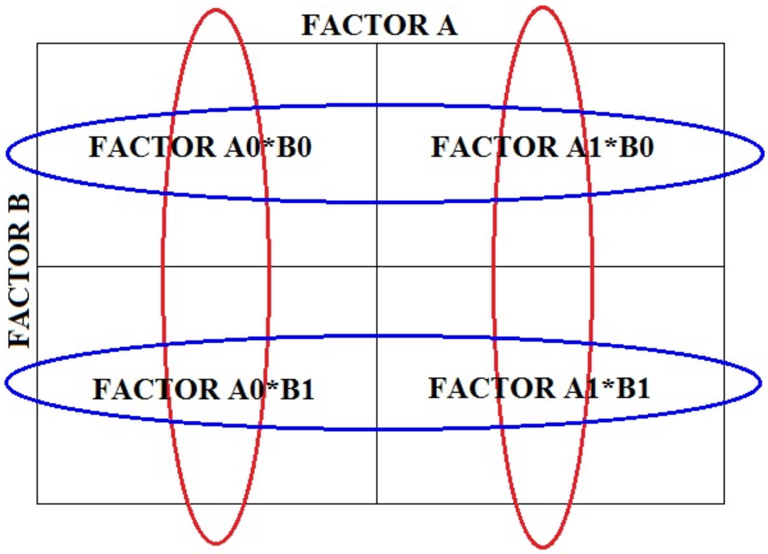


Figure B.7: The groups for a two-way ANOVA on the main effects. In red are the 2 groups considering the columns and in blue the 2 groups considering the rows.

to be compared only the rows or the columns, mixed combinations are not allowed. In Figure B.7 the groups for main effects are shown.

The second values considers the interaction effects, that is the effects of one factor on the others. An $A \cdot B$ interaction is a change in the simple main effect of B over levels of A or the change in the simple main effect of A over levels of B. Here are involved mixed combinations. In Algorithm 3 on page xxii are shown all the steps to perform a two-ways ANOVA.

In the plots in Chapter 4 are shown the improvements of the balance both of main effects and interactions.

B.3.4 Refining the Quintile

The refinement is an iterative process. First all the variables excluded by the model are ranked according to their F-ratios and of those the 25% is inserted one by one in a new model. If after their inclusion, the balance for the current variable is not improving, the square of this variable and then

Algorithm 3 Pseudocode of the Two-Ways ANOVA with 2 binary variables.

Step 1: Consider the main effects, that is rows and columns of the two-ways binary factors A and B
Step 2: Calculate the overall mean m
Step 3: Considering the rows (first factor, A), calculate the means of the 2 rows $(i = 1, 2)$, $m_{a,i}$
Step 4: Considering the rows (first factor, A), calculate the elements of the 2 rows $(i = 1, 2)$, $n_{a,i}$
Step 5: Calculate the between-groups sum of squares for A, $S_{1,a}$
Step 6: Considering each element of the groups $x_{i,j}$ calculate the within-groups sum of squares for A, $S_{2,a}$
Step 7: Calculate the number of groups $k_a = 2$ and the elements in all the groups n
Step 8: Calculate the ratio F_a for factor A using $S_{1,a}$, $S_{2,a}$, k_a and n
Step 9: Repeat the process for the columns (second factor, B), obteining $S_{1,b}$, $S_{2,b}$ and F_b
Step 10: Calculate the between-groups sum of squares, S_{bw}, considering all the 4 groups (2 rows and 2 columns)
Step 11: Calculate $S_{1,a\cdot b} = S_{bw} - S_{1,a}, -S_{1,b}$
Step 12: Calculate the within-groups sum of squares, S_{wi}, considering all the 4 groups (2 rows and 2 columns)
Step 13: Calculate $S_{2,a\cdot b} = S_{wi} - S_{2,a}, -S_{2,b}$
Step 14: Calculate the number of groups $k_{a\cdot b} = 2 \cdot 2$ and the elements in all the groups n
Step 15: Calculate the ratio $F_{a\cdot b}$ for interaction $A \cdot B$ using $S_{1,a\cdot b}$, $S_{2,a\cdot b}$, $k_{a\cdot b}$ and n

the interactions of it with the variables in the model generated in the fitting phase are tried. If none of this possibilieties improves the F-ratio of the analyzed variable, it is discarded. If it improves it is included in the model.

B.4 Outcome Analysis and Machine Learning with GP Analysis

After performing the procedures of the previous steps, the ones relating to these two types of analysis are simple.

As regards the outcome analysis, were carried out a series of regressions performed using the methods provided by Matlab and a series of new stratifications performed using procedures similar to the ones previously described.

As regards the machine learning with gp analysis, as already said, the executions of the genetic programming were carried out using GPLAB.

Appendix C

Details on the Datasets

In this Appendix details on the datasets will be provided.

C.1 List of Diuretics

The diuretics variable was computated by looking in the Mimic II Clinical Database for the following list of drugs:

- acetazolamide (Diamox), dichlorphenamide (Daranide);
- methazolamide(Glauctabs, MZM, Neptazane), torsemide (Demadex), furosemide (Lasix);
- pironolactone (Aldactone), amiloride (Midamor), triamterene (Dyrenium);
- hydrochlorothiazide (HCTZ, HydroDIURIL, Aquazide H, Esidrix, Microzide), metolazone (Mykrox, Zaroxolyn);
- methyclothiazide (Enduron, Aquatensen), chlorothiazide (Diuril), indapamide (Lozol);
- bendroflumethiazide (Naturetin), polythiazide (Renese), hydroflumethiazide (Saluron), chlorthalidone (Thalitone).

This list has been suggested by the medical experts.

C.2 List of Fluids

The fluids inputs variable has be computated looking in the Mimic II Clinical Database for the following list of items:

- 106 - Lactated Ringers, 107 - .9% Normal Saline, 130 - D5/.45NS, 131 - D5/.45NS 10000.0ml;
- 134 - .9% Normal Saline 1000.0ml, 142 - Lactated Ringers 1000.0ml, 151 - 45% Normal Saline 1000.0ml, 152 - D5/.45NS 1000.0ml;
- 154 - D5NS, 165 - D5W 1000.0ml, 180 - .45% Normal Saline, 187 - .9% Normal Saline 500.0ml;
- 214 - D5 Normal Saline, 219 - D5RL 1000.0ml, 249 - .9% Normal Saline 250.0ml, 297 - D5NS 1000.0ml;
- 299 - D5 Normal Saline 1000.0ml, 309 - .9% Normal Saline 100.0ml, 615 - D5/.45NS 2000.0ml, 631 - .9% Normal Saline 2000.0ml.

The number next to each fluid is the corresponding identifier in the database.

C.3 Variables Descriptive Statistics

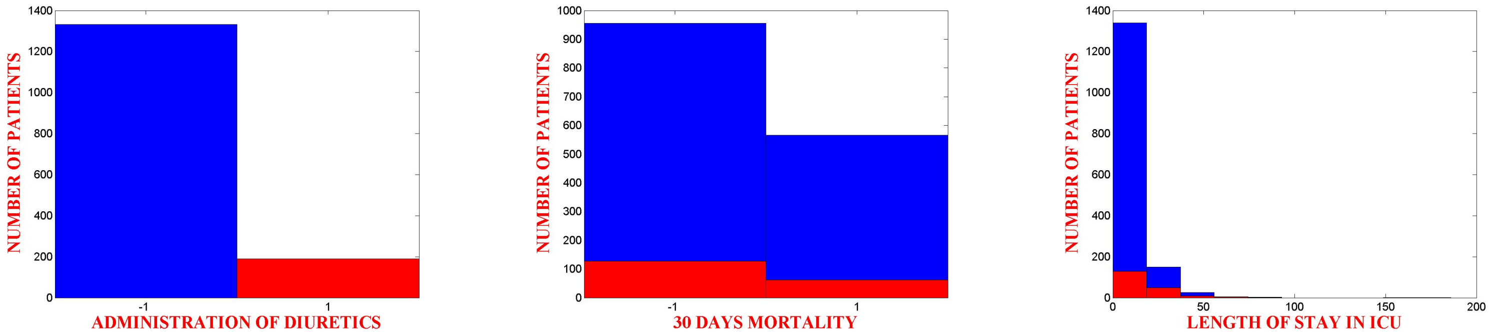
In this Section histograms of the values of all the variables are provided.

Figure C.1 on page xxvii show the histograms for diuretics, mortality and length of stay.

Figures C.2 on page xxviii and C.2 on page xxviii show the histograms for gender, race, use of vasopressor and mechanical ventilation.

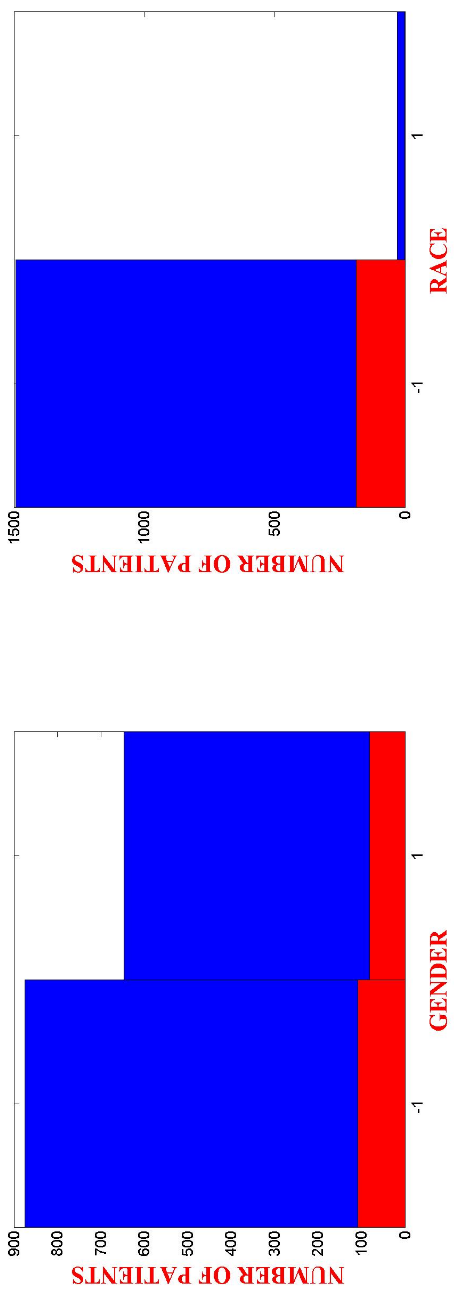
Figures C.4 on page xxx and C.5 on page xxxi show the histograms for the 9 Elixhauser parameters.

Figures C.6 on page xxxii, C.7 on page xxxiii, C.8 on page xxxiv, C.9 on page xxxv, C.10 on page xxxvi, C.11 on page xxxvii, C.12 on page xxxviii show the histograms of all the numeric variables.



a: 189 patients got diuretics and 133 didn't. *b: 566 of 1522 patients died within 30 days.* *c: Length of Stay is centered on 7.4 days.*

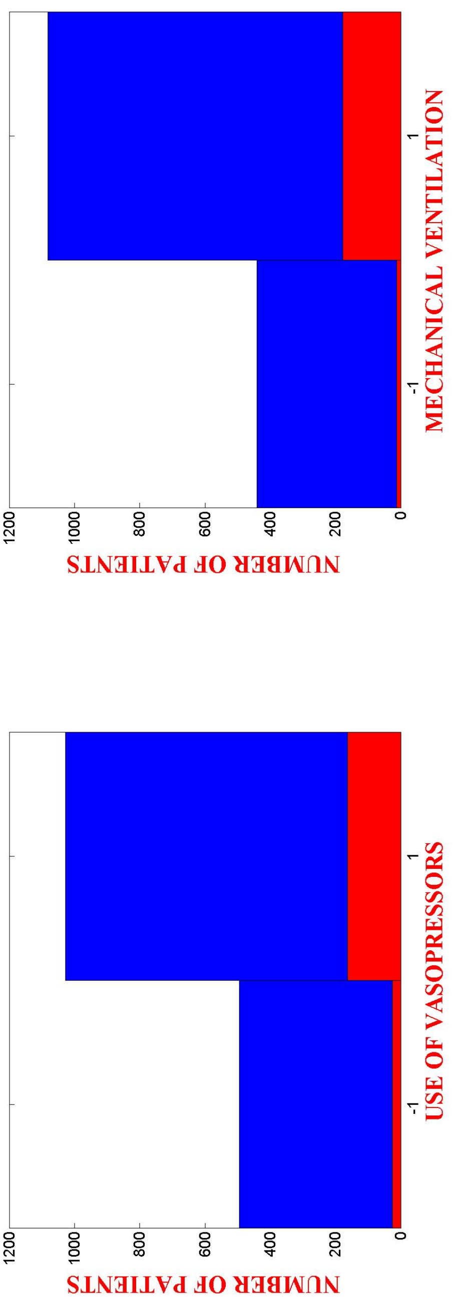
Figure C.1: Histograms of diuretics, mortality and length of stay.



a: 647 of 1522 patients are female and 875 male.

b: 1492 of 1522 patients are not white.

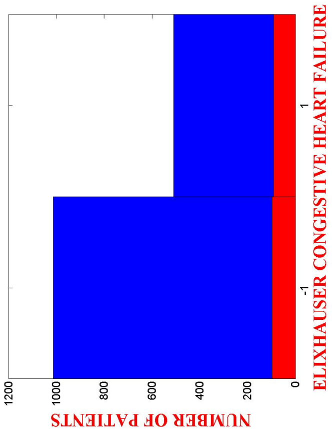
Figure C.2: Histograms of gender and race.



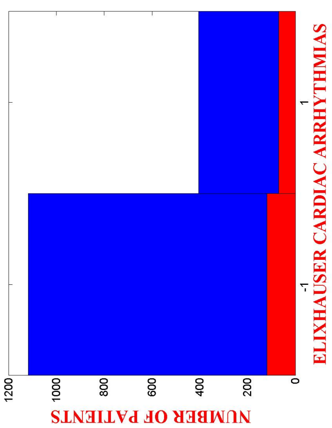
a: 1028 of 1522 patients are on vasopressors.

b: 1081 of 1522 patients are on mechanical ventilation

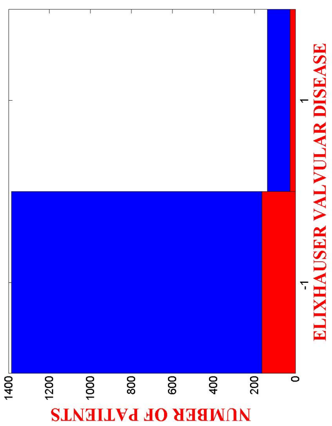
Figure C.3: Histograms of use of vasopressor and mechanical ventilation.



a: 509 patients had congestive heart failure.

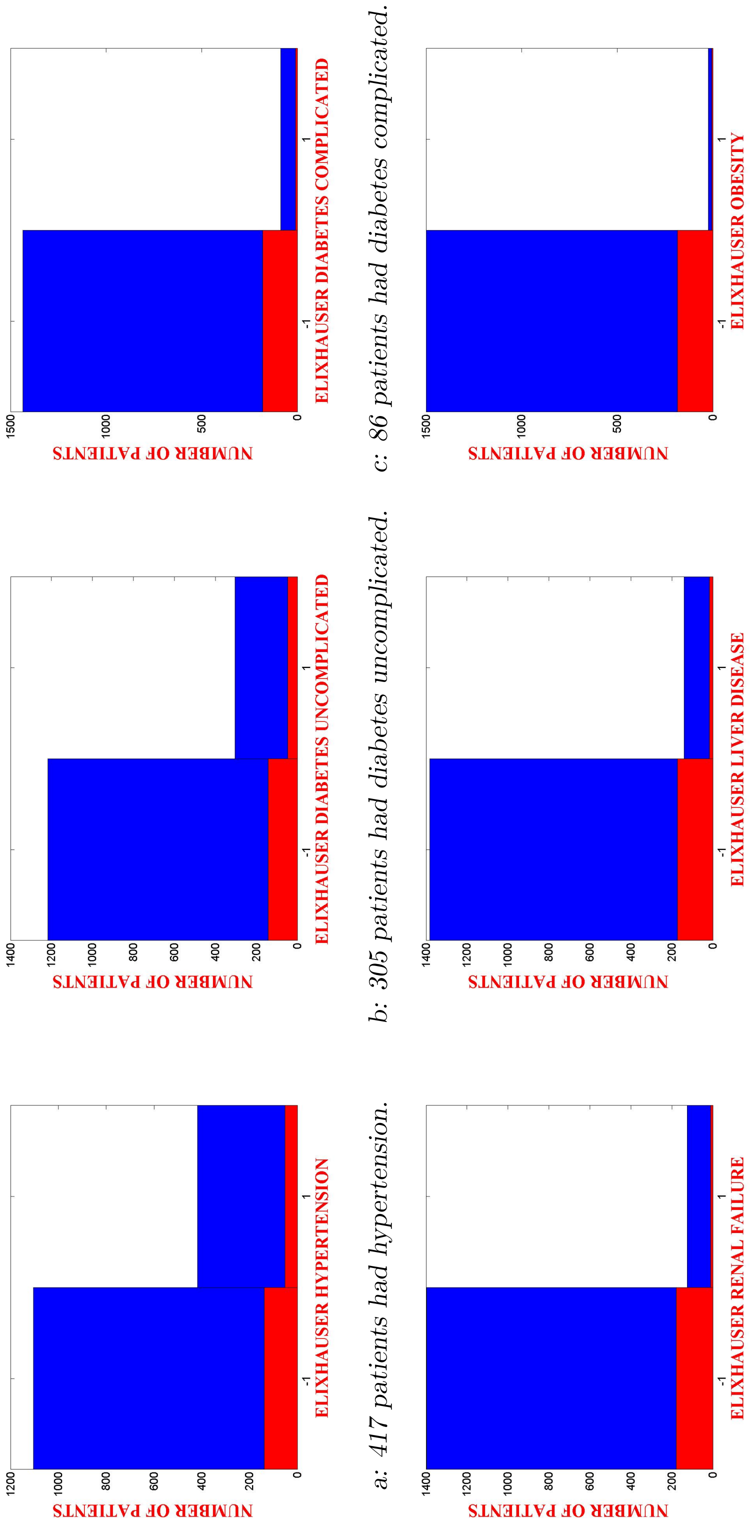


b: 405 patients had cardiac arrhythmias.



c: 136 patients had valvular disease.

Figure C.4: Histograms of the Elixahuser parameters, part 1.



a: 417 patients had hypertension.

b: 305 patients had diabetes uncomplicated.

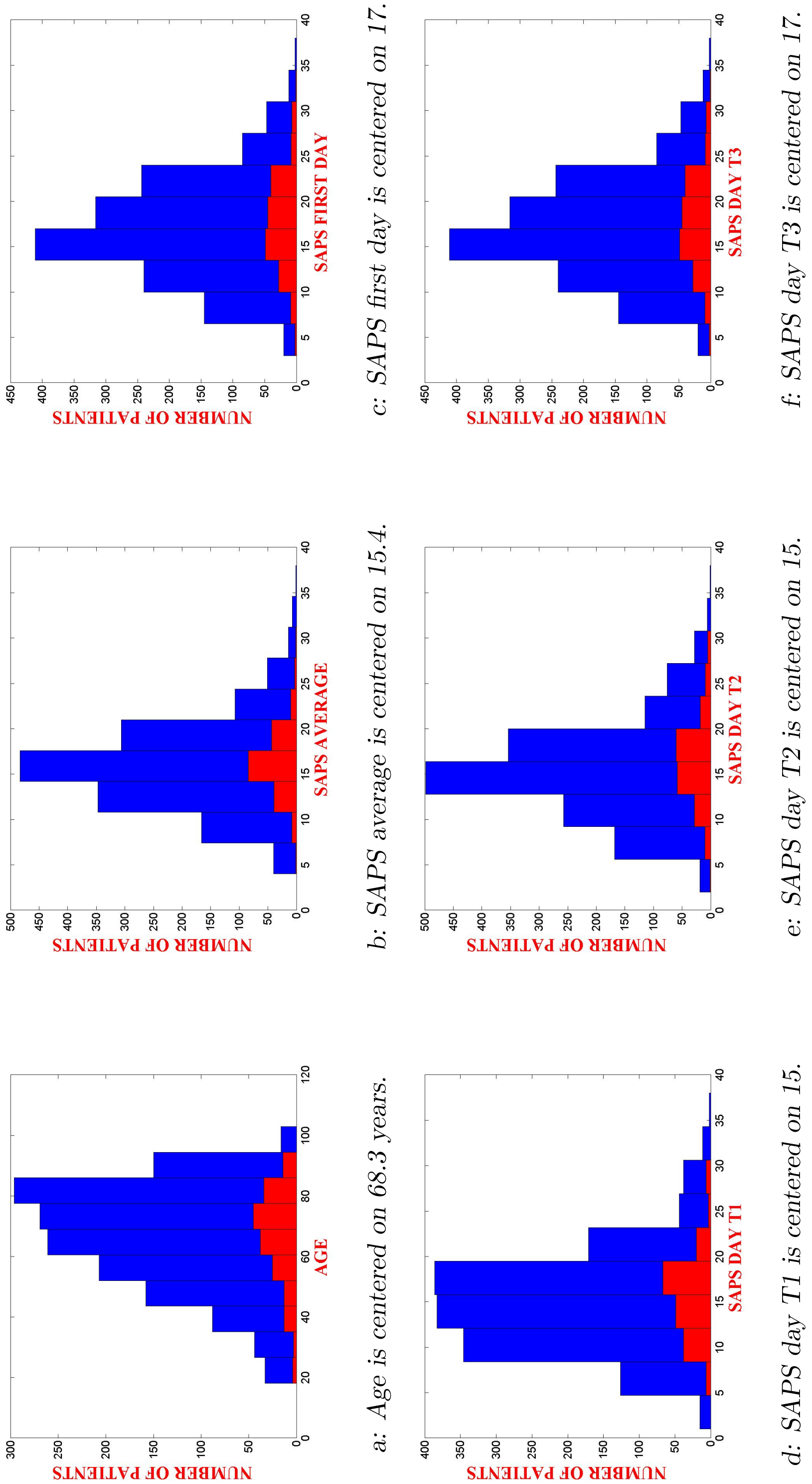
c: 86 patients had diabetes complicated.

d: 125 patients had renal failure.

e: 141 patients had liver disease..

f: 23 patients had obesity.

Figure C.5: Histograms of the Elixahuser parameters, part 2.



a: Age is centered on 68.3 years.

b: SAPS average is centered on 15.4.

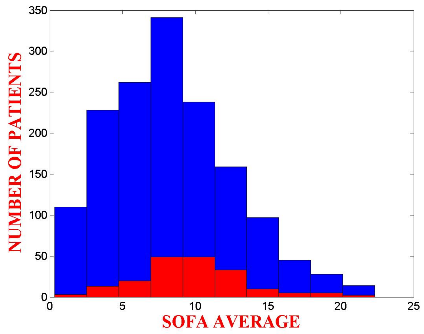
c: SAPS first day is centered on 17.

d: SAPS day T1 is centered on 15.

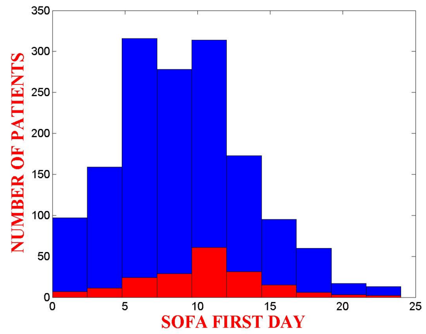
e: SAPS day T2 is centered on 15.

f: SAPS day T3 is centered on 17.

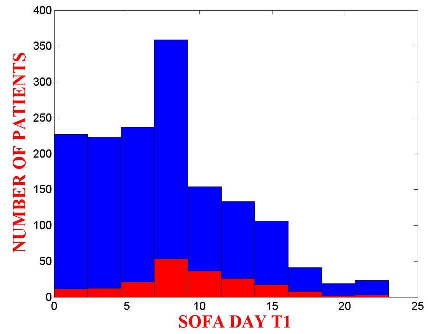
Figure C.6: Histograms of the numeric variables, part 1



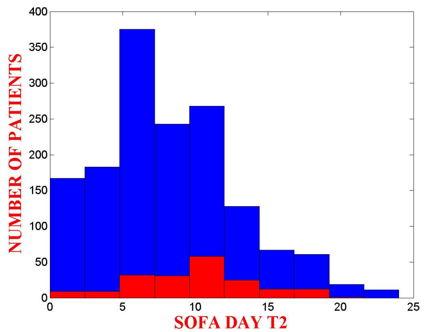
a: *SOFA average is centered on 7.6.*



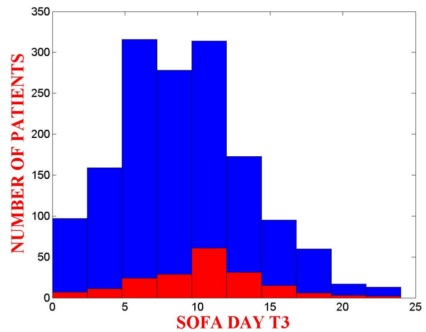
b: *SOFA first day is centered on 9.*



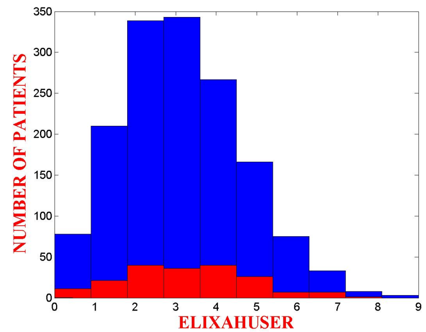
c: *SOFA day T1 is centered on 7.*



d: *SOFA day T2 is centered on 7.4.*

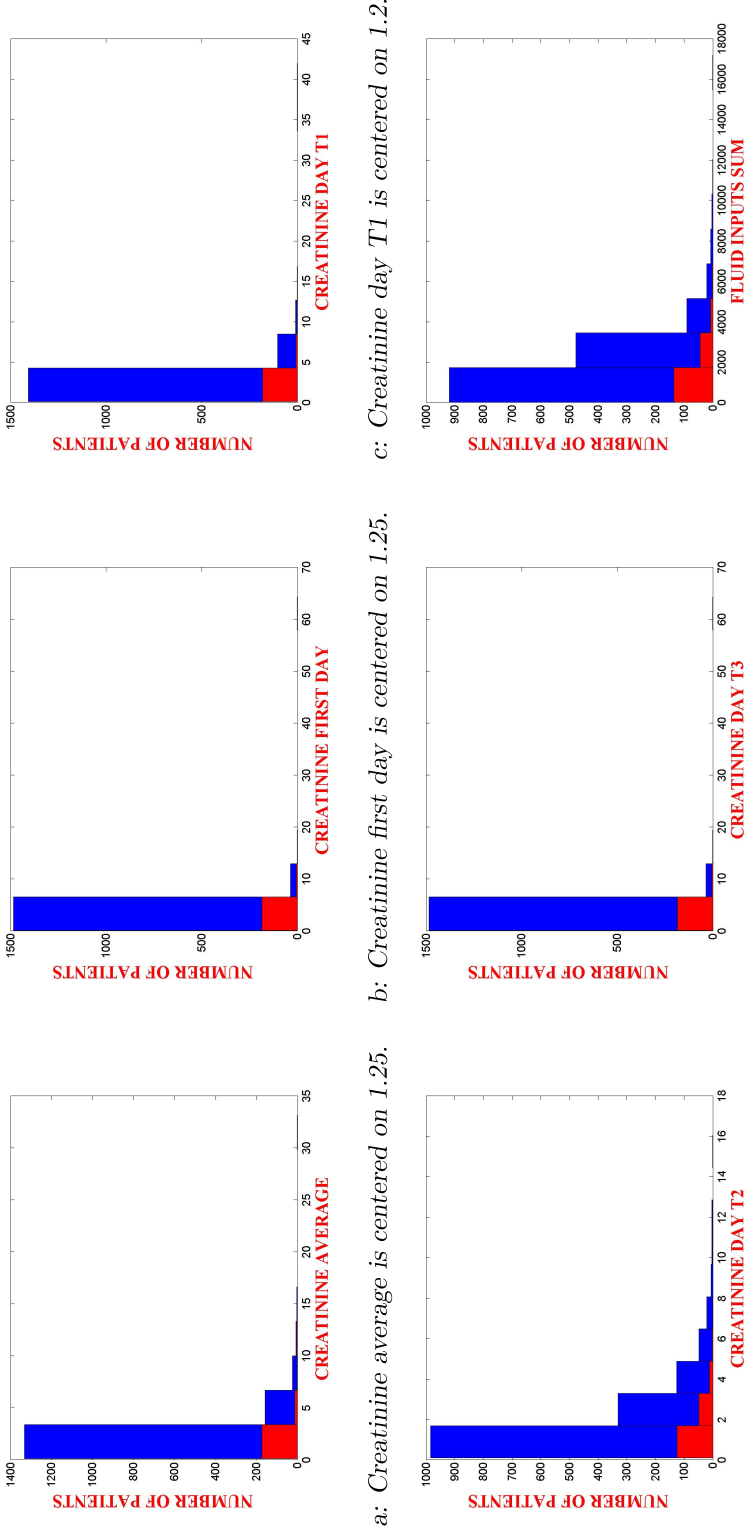


e: *SOFA day T3 is centered on 9.*



f: *Elixahuser is centered on 3.*

Figure C.7: Histograms of the numeric variables, part 2



a: Creatinine average is centered on 1.25.

b: Creatinine first day is centered on 1.25.

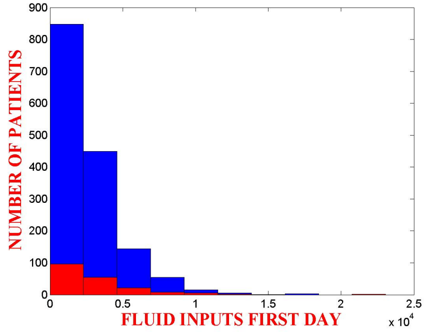
c: Creatinine day T1 is centered on 1.2.

d: Creatinine day T2 is centered on 1.2.

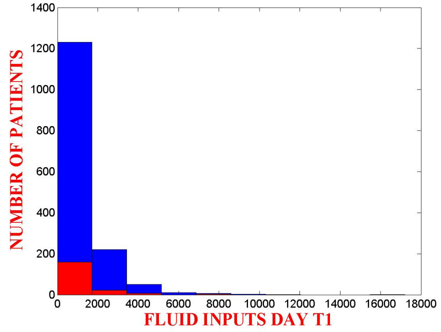
e: Creatinine day T3 is centered on 1.25.

f: Fluids inputs sum is centered on 1.4 liters.

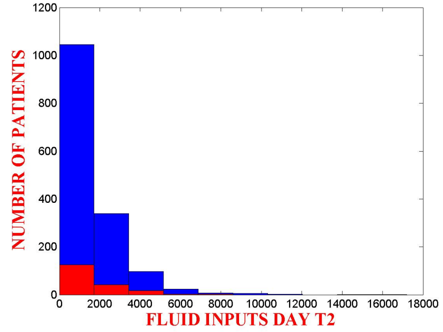
Figure C.8: Histograms of the numeric variables, part 3



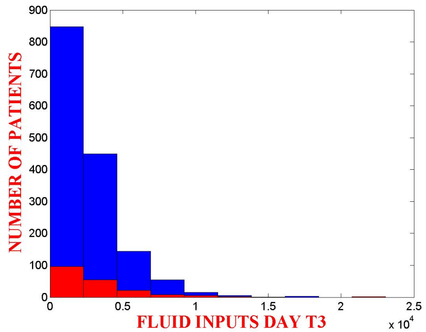
a: *Fluids inputs first day is centered on 2 liters.*



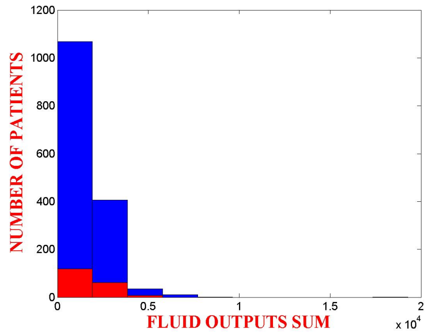
b: *Fluids inputs day T1 is centered on 0.7 liters.*



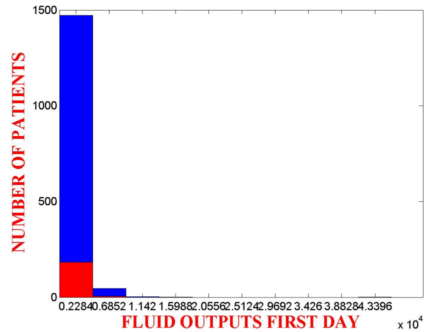
c: *Fluids inputs day T2 is centered on 1.0 liters.*



d: *Fluids inputs day T3 is centered on 2.0 liters.*

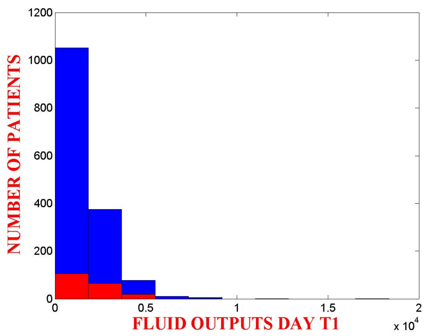


e: *Fluids outputs sum is centered on 1.3 liters.*

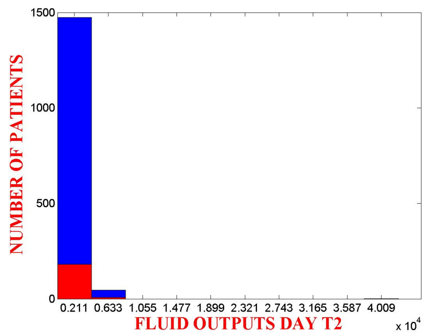


f: *Fluids outputs first day is centered on 1.2 liters.*

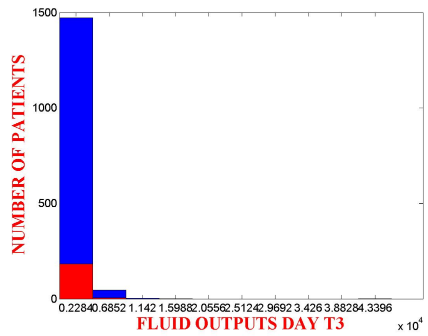
Figure C.9: Histograms of the numeric variables, part 4



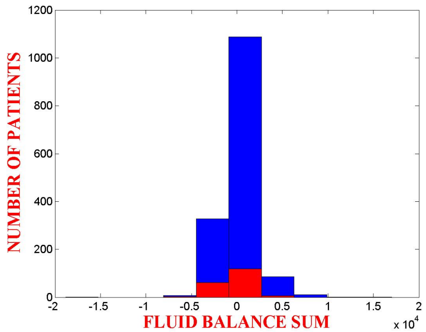
a: *Fluids outputs day T1 is centered on 1.1 liters.*



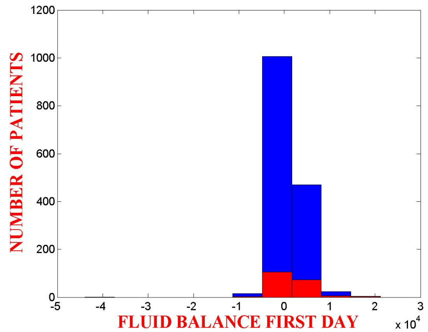
b: *Fluids outputs day T2 is centered on 1.1 liters.*



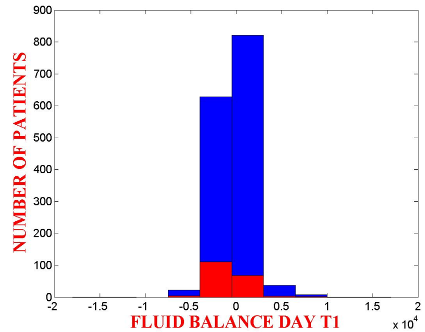
c: *Fluids outputs day T3 is centered on 1.2 liters.*



d: *Fluids balance sum is centered on 0.05 liters.*

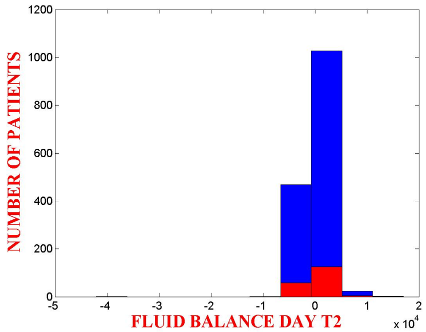


e: *Fluids balance first day is centered on 0.67 liters.*

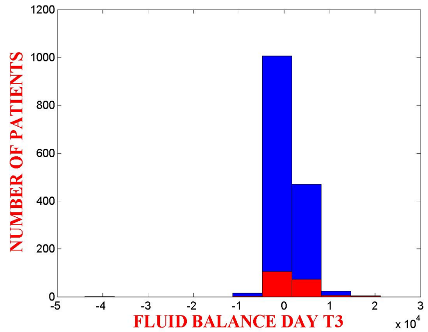


f: *Fluids balance day T1 is centered on -0.25 liters.*

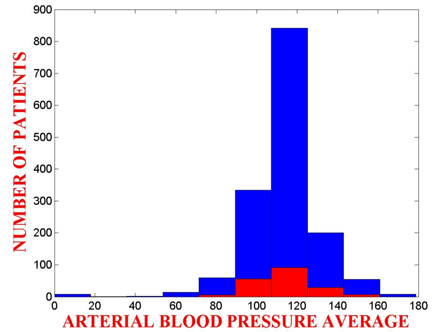
Figure C.10: Histograms of the numeric variables, part 5



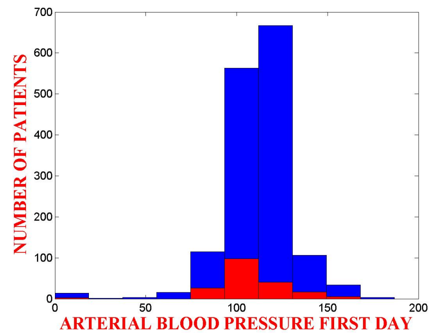
a: *Fluids balance day T2 is centered on -0.002 liters.*



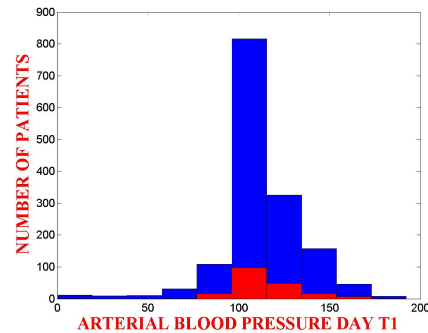
b: *Fluids balance day T3 is centered on 0.67 liters.*



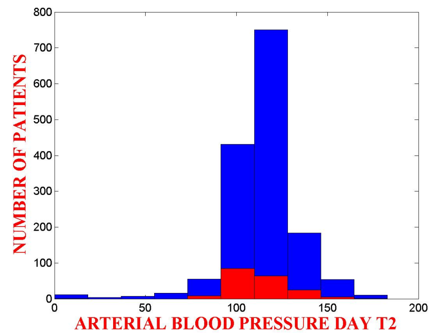
c: *Arterial blood pressure average is centered on 114.9.*



d: *Arterial blood pressure first day is centered on 114.1.*

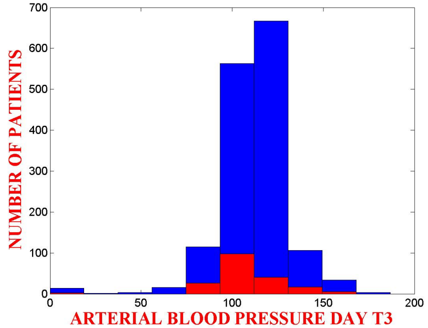


e: *Arterial blood pressure day T1 is centered on 114.9.*

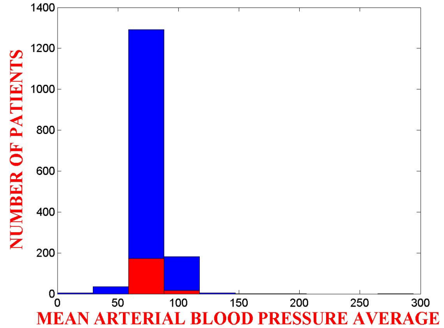


f: *Arterial blood pressure day T2 is centered on 114.9.*

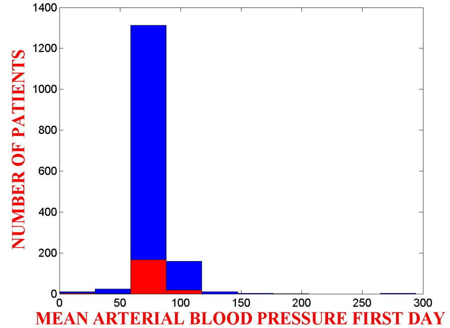
Figure C.11: Histograms of the numeric variables, part 6



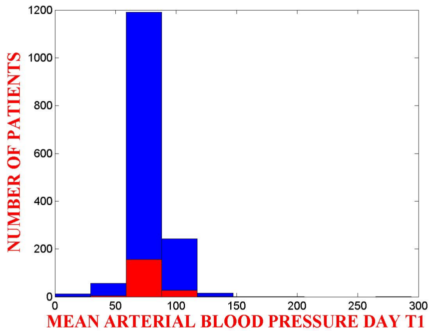
a: Arterial blood pressure day T3 is centered on 114.1.



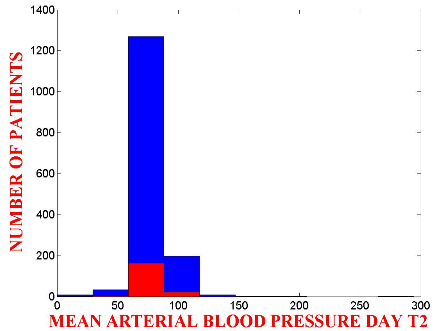
b: Mean arterial blood pressure average is centered on 79.2.



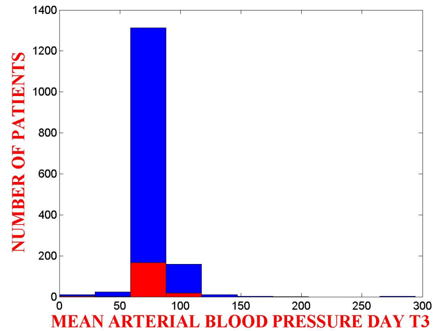
c: Mean arterial blood pressure first day is centered on 79.2.



d: Mean arterial blood pressure day T1 is centered on 79.2.



e: Mean arterial blood pressure day T2 is centered on 79.2.



f: Mean arterial blood pressure day T3 is centered on 79.2.

Figure C.12: Histograms of the numeric variables, part 7

C.4 Timeline Values Discussion

In this Section an overview of the correlations of the values for the variables with timelines at 8 time points will be provided. The final decisions on how to prepare timeline variables is in Chapter 2.

Here the earlier data that have been collected which informed the final decisions are described. Follows the list of the studied time points:

Timepoint 1: Diuretics average over D^+ patients (189 of 1,522 patients);

Timepoint 2: Diuretics average over D^+ patients as fraction of length of stay;

Timepoint 3: First fluids balance minimum;

Timepoint 4: Second fluids balance minimum;

Timepoint 5: Stop of vasopressors average on 1,028 of 1,522 patients;

Timepoint 6: Stop of vasopressors average on 1,028 of 1,522 patients as fraction of length of stay;

Timepoint 7: First blood pressure minimum;

Timepoint 8: Second blood pressure minimum.

These results showed that most of the defined time points were correlated and not introduced new information regarding diuretics to the dataset. So, it was decided to adopt the time points defined in Chapter 2

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1	0.75	0.70	0.75	0.83	0.75	0.72	0.76
T2	0.75	1	0.86	0.88	0.79	0.87	0.83	0.85
T3	0.70	0.86	1	0.89	0.75	0.82	0.86	0.84
T4	0.75	0.88	0.89	1	0.78	0.86	0.85	0.89
T5	0.83	0.79	0.75	0.78	1	0.90	0.77	0.79
T6	0.75	0.87	0.82	0.86	0.90	1	0.81	0.84
T7	0.72	0.83	0.86	0.85	0.77	0.81	1	0.90
T8	0.76	0.85	0.84	0.89	0.79	0.84	0.90	1

Table C.1: The correlations between 8 time points for the saps variable.

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1	0.74	0.70	0.75	0.80	0.72	0.72	0.76
T2	0.74	1	0.87	0.89	0.79	0.89	0.86	0.88
T3	0.70	0.87	1	0.91	0.77	0.84	0.89	0.87
T4	0.75	0.89	0.91	1	0.79	0.87	0.87	0.91
T5	0.80	0.79	0.77	0.79	1	0.90	0.79	0.80
T6	0.72	0.89	0.84	0.87	0.90	1	0.85	0.86
T7	0.72	0.86	0.89	0.87	0.79	0.85	1	0.93
T8	0.76	0.88	0.87	0.91	0.80	0.86	0.93	1

Table C.2: The correlations between 8 time points for the sofa variable.

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1	0.89	0.37	0.38	0.47	0.45	0.46	0.47
T2	0.89	1	0.43	0.42	0.48	0.49	0.49	0.50
T3	0.37	0.43	1	0.99	0.81	0.82	0.82	0.82
T4	0.38	0.42	0.99	1	0.81	0.82	0.82	0.82
T5	0.47	0.48	0.81	0.81	1	0.99	0.97	0.97
T6	0.45	0.49	0.82	0.82	0.99	1	0.98	0.98
T7	0.46	0.49	0.82	0.82	0.97	0.98	1	0.99
T8	0.47	0.50	0.82	0.82	0.97	0.98	0.99	1

Table C.3: The correlations between 8 time points for the creatinine variable.

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1	0.66	0.58	0.61	0.68	0.60	0.61	0.63
T2	0.66	1	0.67	0.68	0.55	0.62	0.65	0.66
T3	0.58	0.67	1	0.93	0.62	0.68	0.80	0.78
T4	0.61	0.68	0.93	1	0.64	0.71	0.78	0.78
T5	0.68	0.55	0.62	0.64	1	0.87	0.66	0.68
T6	0.60	0.62	0.68	0.71	0.87	1	0.66	0.68
T7	0.61	0.65	0.80	0.78	0.66	0.66	1	0.92
T8	0.63	0.66	0.78	0.78	0.68	0.68	0.92	1

Table C.4: The correlations between 8 time points for the fluids inputs variable.

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1	0.47	0.38	0.40	0.50	0.35	0.37	0.48
T2	0.47	1	0.55	0.57	0.43	0.50	0.36	0.42
T3	0.38	0.55	1	0.72	0.45	0.49	0.56	0.54
T4	0.40	0.57	0.72	1	0.45	0.52	0.55	0.55
T5	0.50	0.43	0.45	0.45	1	0.80	0.60	0.74
T6	0.35	0.50	0.49	0.52	0.80	1	0.60	0.66
T7	0.37	0.36	0.56	0.55	0.60	0.60	1	0.76
T8	0.48	0.42	0.54	0.55	0.74	0.66	0.76	1

Table C.5: The correlations between 8 time points for the fluids outputs variable.

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1	0.35	0.30	0.36	0.46	0.40	0.29	0.35
T2	0.35	1	0.39	0.42	0.38	0.42	0.32	0.34
T3	0.30	0.39	1	0.68	0.36	0.42	0.55	0.47
T4	0.36	0.42	0.68	1	0.41	0.40	0.52	0.53
T5	0.46	0.38	0.36	0.41	1	0.80	0.42	0.45
T6	0.40	0.42	0.42	0.40	0.80	1	0.35	0.39
T7	0.29	0.32	0.55	0.52	0.42	0.35	1	0.68
T8	0.35	0.34	0.47	0.53	0.45	0.39	0.68	1

Table C.6: The correlations between 8 time points for the fluids balance variable.

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1	0.63	0.54	0.62	0.79	0.71	0.62	0.63
T2	0.63	1	0.64	0.65	0.60	0.64	0.64	0.65
T3	0.54	0.64	1	0.85	0.57	0.62	0.71	0.73
T4	0.62	0.65	0.85	1	0.62	0.66	0.68	0.73
T5	0.79	0.60	0.57	0.62	1	0.91	0.59	0.60
T6	0.71	0.64	0.62	0.66	0.91	1	0.62	0.62
T7	0.62	0.64	0.71	0.68	0.59	0.62	1	0.83
T8	0.63	0.65	0.73	0.73	0.60	0.62	0.83	1

Table C.7: The correlations between 8 time points for the vasopressors amounts variable.

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1	0.55	0.53	0.60	0.68	0.59	0.58	0.60
T2	0.55	1	0.64	0.59	0.59	0.62	0.63	0.60
T3	0.53	0.64	1	0.82	0.59	0.63	0.71	0.65
T4	0.60	0.59	0.82	1	0.60	0.65	0.67	0.73
T5	0.68	0.59	0.59	0.60	1	0.88	0.60	0.60
T6	0.59	0.62	0.63	0.65	0.88	1	0.62	0.62
T7	0.58	0.63	0.71	0.67	0.60	0.62	1	0.84
T8	0.60	0.60	0.65	0.73	0.60	0.62	0.84	1

Table C.8: The correlations between 8 time points for the blood pressure variable.

C.5 Dataset Correlations

Tables C.9 and C.10 on page xliv show an overview of the correlations between the variables of the study at an earlier point at time. Table C.11 on page xlv provide full description of the abbreviations. This shows that all the parameters are sensible and that the weight of the possible outliers present in the database has been mitigated. In red are the correletion values which are significant.

	Di	Sa1	Sa2	So1	So2	Eli	Cr1	Cr2	Bp1	Bp2	MO	LS
Di	1	-0.1	0,08	0,01	0,14	0,03	-0,0	-0,0	-0.0	-0.0	-0.0	0.37
Sa1	-0.1	1	0,48	0,62	0,32	0,03	0.16	0,16	-0.1	-0.0	0.4	-0.2
Sa2	0,08	0,48	1	0,50	0,67	0,04	0,18	0,21	-0.0	-0.0	0.37	0.14
So1	0,01	0,62	0,50	1	0,69	0,02	0,29	0,29	-0.2	-0.0	0.56	-0.0
So2	0,14	0,32	0,67	0,69	1	0,04	0,25	0,30	-0.1	-0.0	0.38	0.16
Eli	0,03	0,03	0,04	0,02	0,04	1	0,06	0,08	-0.0	-0.0	0.07	-0.0
Cr1	-0,0	0,16	0,18	0,29	0,25	0,06	1	0.74	-0.0	-0.0	0.16	-0.0
Cr2	-0,0	0,16	0,21	0,29	0,30	0,08	0.74	1	-0.0	-0.0	0.16	-0.0
Bp1	-0.0	-0.1	-0.0	-0.2	-0.1	-0.0	-0.0	-0.0	1	0.40	-0.1	-0.0
Bp2	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.40	1	-0.0	0.00
MO	-0,0	0.4	0.37	0.56	0.38	0.07	0.16	0.16	-0.1	-0.0	1	-0.0
LS	0.37	-0.2	0.14	-0.0	0.16	-0.0	-0.0	-0.0	-0.0	0.00	-0.0	1

Table C.9: Correlations between a subset of the variables of the final dataset. In red the values with high correlation.

	Fi1	Fi2	Fo1	Fo2	Fb1	Fb2	Fb3	Va1	Va2	Ve	MO	LS
Fi1	1	0,30	0,06	0,04	0.70	0,33	0,11	0,23	0.06	0,01	0,11	-0,1
Fi2	0,30	1	0,03	0,03	0.20	0,18	0,16	0,10	0,09	0,07	0,01	0,01
Fo1	0,06	0,03	1	0,38	-0.6	0,09	0,12	-0.1	-0,0	0,01	-0,3	0,03
Fo2	0,04	0,03	0,38	1	-0.2	0,01	0,04	-0,0	0,00	0,00	-0,1	0,08
Fb1	0,70	0,20	-0,6	-0,2	1	0,18	-0,0	0,28	0,08	-0,0	0,37	-0,1
Fb2	0,33	0,18	0,09	0,01	0.18	1	0,23	0.25	0,10	0,29	0,14	0,15
Fb3	0,11	0,16	0,12	0,04	-0.0	0,23	1	0,03	0,03	0,08	-0,0	0,02
Va1	0.23	0,10	-0,1	-0,0	0.28	0,25	0,03	1	0,44	0,20	0,33	-0,0
Va2	0,06	0,09	-0,0	0,00	0.08	0,10	0,03	0,44	1	0,16	0,15	0,02
Ve	0.01	0.07	0,01	0,00	-0.0	0,29	0.08	0,20	0.16	1	0,22	0,35
MO	0.11	0,01	-0,3	-0,1	0.37	0,14	-0,0	0,33	0,15	0,22	1	-0,0
LS	-0,1	0,01	0.03	0,08	-0.1	0,15	0,02	-0,0	0,02	0,35	-0,0	1

Table C.10: Correlations between a subset of the variables of the final dataset. In red the values with high correlation.

Abbreviation	Description
Di	Abministration of Diuretics
Sa1	SAPS at the first day in ICU
Sa2	SAPS sum during the stay in ICU
So1	SOFA at the first day in ICU
So2	SOFA sum during the stay in ICU
Eli	Elixahuser Score
Cr1	Creatinine at the first day in ICU
Cr2	Creatinine sum during the stay in ICU
Bp1	Arterial Blood Pressure at the first day in ICU
Bp2	Arterial Blood Pressure sum during the stay in ICU
Fi1	Fluids Inputs at the first day in ICU
Fi2	Fluids Inputs sum during the stay in ICU
Fo1	Fluids Outputs at the first day in ICU
Fo2	Fluids Outputs sum during the stay in ICU
Fb1	Balance $Input - Outputs$ sum during the stay in ICU
Fb2	Balance sum from $TotalBalanceEvents$
Fb3	Balance at the first day in ICU from $TotalBalanceEvents$
Va1	Amount of Vasopressors at the first day in ICU
Va2	Amount of Vasopressors sum during the stay in ICU
Ve	Mechanical Ventilation
MO	Mortality
LS	Length of Stay

Table C.11: Abbrevation used in the correlations tables.

C.6 Experts Datasets

Except for the datasets described in Chapter 3, the propensity analysis was performed on two more lists provided by medical experts[1]. In this Section the results on this two datasets will be presented.

List of the variables chosen by the doctors follow:

- **Experts list 1:** Chosen variables:
  1. Age when admitted in the ICU (x_2)
  2. Race (white vs not white) (x_4)
  3. Elixhauser overall (x_15)
  4. Elixhauser binary (selected 9 fields) ($x_16 \rightarrow x24$)
  5. Creatinine mean of values during the first day (x_26)
  6. Fluids inputs sum of values during the first day (x_31)
  7. Fluids outputs sum of values during the first day (x_36)
  8. Fluids balance sum of values during the first day (x_41)
  9. Use of vasopressors in the ICU (x_45)
  10. Mechanical ventilation in the ICU (x_46)
  11. Arterial bp mean of values during the first day (x_48)
- **Experts list 2:** Chosen variables:
  1. Age when admitted in the ICU (x_2)
  2. Race (white vs not white) (x_4)
  3. Elixhauser overall (x_15)
  4. Elixhauser binary (selected 9 fields) ($x_16 \rightarrow x24$)
  5. Creatinine mean of values during the first day (x_26)
  6. Fluids inputs sum of values during the first day (x_31)
  7. Fluids outputs sum of values during the first day (x_36)
  8. Fluids balance sum of values during the first day (x_41)
  9. Use of vasopressors in the ICU (x_45)
  10. Mechanical ventilation in the ICU (x_46)
  11. Arterial bp mean of values during the first day (x_48)

[1] The work has been performed with the support of medical experts, including Dr. Leo Celi, MD Critical Care Physician - Boston, MA and John Danziger, MD Department of Medicine, Division of Nephrology, Beth Israel Deaconess Medical Center - Boston, MA.

12. Creatinine mean of values during day T1 (x_27)
13. Fluids inputs sum of values during day T1 (x_32)
14. Fluids outputs sum of values during day T1 (x_37)
15. Fluids balance sum of values during day T1 (x_42)

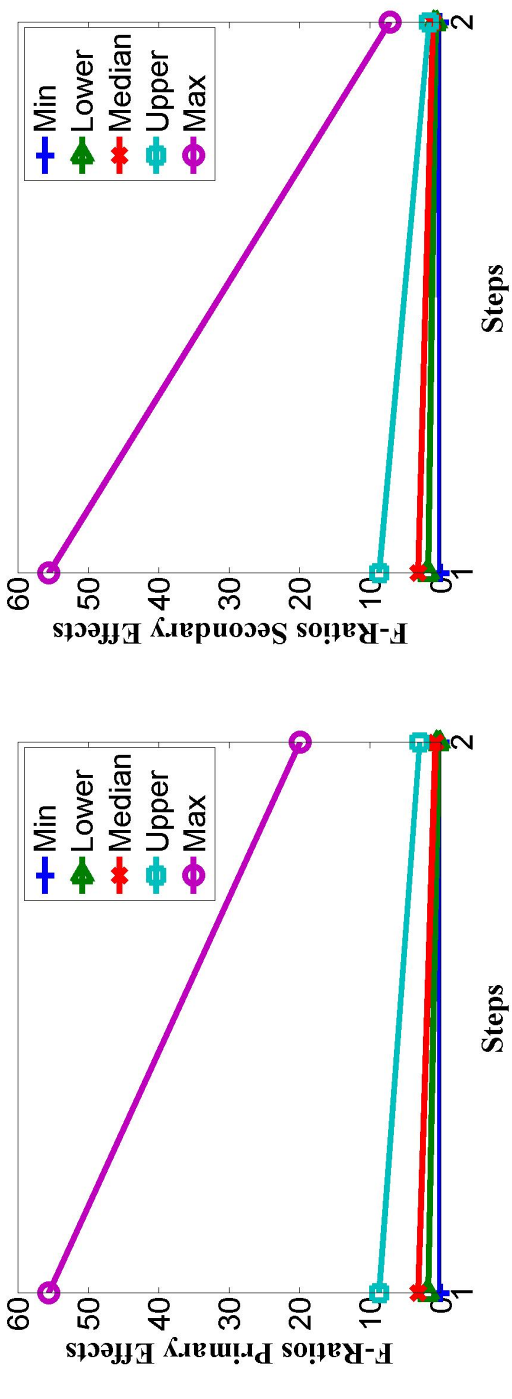
In Figures C.13 on page xlviii and C.14 on page xlix the improvements in the balance after subclassifiction with the list of variables provided above are shown:

Finally, a comparison of the results between these 2 new datasets is now provided:

- **Experts list 1:** These quintileswere created with the help of medical experts. They seems to be more balanced then the ones created with the stepwise method, because the propensity number calculated in this way was less predictive. The results for this dataset are shown in Table C.12 on page l.

  Quintile 1 is unbalanced and it should not be considered for the analysis. The length of stay in ICU appears to be longer for the patients who got diuretics in this case too, while the chances of survival are better for patients who did not get diuretics belonging to quintile 3 and for patients who got diuretics belonging to quintile 5.

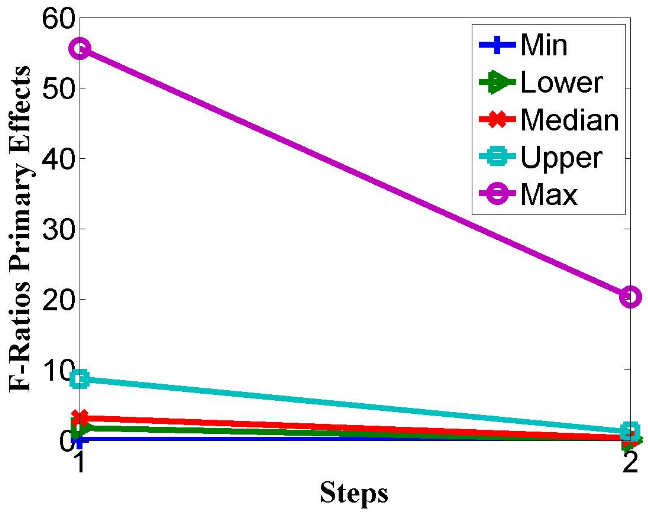
- **Experts list 2:** These quintileswere created with the help of medical experts too and they are similar to, even though not the same, the previous one. In fact in this case the accurancy of the model seems to be improved.



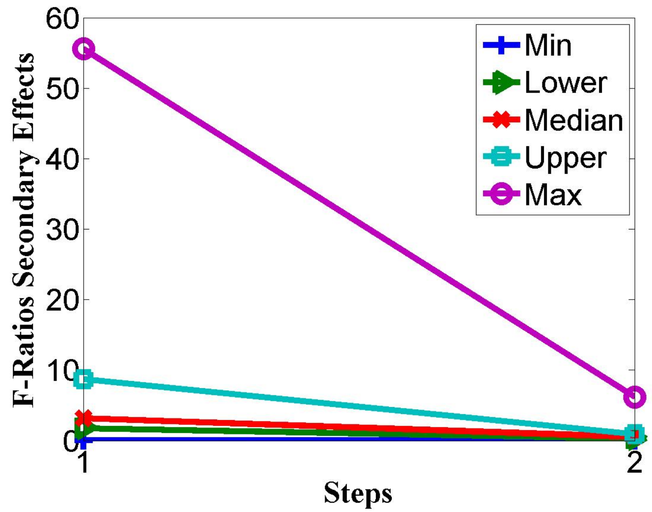
a: Primary Effects.

b: Secondary Effects

Figure C.13: The F-statistics on the experts list 1 dataset on the primary and secondary effects. The values in abscissa 1 refers to the balance on the original dataset, while on abscissa 2 there are the balance after the propensity score method.



a: Primary Effects.



b: Secondary Effects

Figure C.14: The F-statistics on the experts list 2 dataset on the primary and secondary effects. The values in abscissa 1 refers to the balance on the original dataset, while on abscissa 2 there are the balance after the propensity score method.

Quintile 1 $PS \in [0.00; 0.02]$	Diuretics given	Diuretics not given
Number of patients	4	300
Deaths	0%	26%
Mean length of stay	4 days	2.3 days
Quintile 2 $PS \in [0.02; 0.08]$	Diuretics given	Diuretics not given
Number of patients	18	286
Deaths	44%	26%
Mean length of stay	12 days	5 days
Quintile 3 $PS \in [0.08; 0.13]$	Diuretics given	Diuretics not given
Number of patients	34	270
Deaths	41%	48%
Mean length of stay	10.9 days	7 days
Quintile 4 $PS \in [0.13; 0.19]$	Diuretics given	Diuretics not given
Number of patients	46	258
Deaths	39%	43%
Mean length of stay	12.4 days	9.9 days
Quintile 5 $PS \in [0.19; 0.65]$	Diuretics given	Diuretics not given
Number of patients	86	218
Deaths	34%	47%
Mean length of stay	19.6 days	8.2 days

Table C.12: Results on experts list 1 dataset.

Quintile 1 $PS \in [0.00; 0.02]$	Diuretics given	Diuretics not given
Number of patients	2	302
Deaths	0%	29%
Mean length of stay	3.5 days	2.2 days
Quintile 2 $PS \in [0.02; 0.06]$	Diuretics given	Diuretics not given
Number of patients	15	289
Deaths	26%	32%
Mean length of stay	14.2 days	5 days
Quintile 3 $PS \in [0.06; 0.12]$	Diuretics given	Diuretics not given
Number of patients	28	276
Deaths	42%	45%
Mean length of stay	12.4 days	7.2 days
Quintile 4 $PS \in [0.12; 0.19]$	Diuretics given	Diuretics not given
Number of patients	46	258
Deaths	32%	46%
Mean length of stay	11.5 days	8.9 days
Quintile 5 $PS \in [0.19; 0.77]$	Diuretics given	Diuretics not given
Number of patients	97	207
Deaths	42%	37%
Mean length of stay	18.2 days	8.3 days

Table C.13: Results on experts list 2 dataset.

Appendix D

Statistical Methods

In this Appendix will be provided some details on the used statistical methods.

D.1 Basic Stuff on Calculating a Propensity Score

It is essential to realize that the outcome variable is not used in this step.

D.1.1 Using the Propensity Score

There are different ways of using the propensity score even if, regardless of the technique, it is always calculated in the same way. However, once the propensity score is calculated, its application is different, and this will now be described. The following contents are drawn from[13].

The three most common analytical techniques based on the propensity score are matching, stratification and regression adjustment. In the study described in Chapter 3, the stratification approach have been followed.

- **Matching by Propensity Scores:** Matching is a technique for adjusting baseline characteristics. Control subjects are matched with treatment subjects on important baseline characteristics, which need to be controlled for (potential confounders). However, an important drawback is the difficulty in finding close matches for all important confounders. The more confounders that require matching, the harder it is to find suitable patients in each group, with a corresponding reduction in sample size. As already noted, propensity scoring summarizes all measured confounders in a single score. So, using the propensity score requires the matching of only one (composite) factor and offers

greater ease of use. This is one of the great advantages of this statistical technique.

- **Stratification by Propensity Scores:** Stratification is another commonly used technique in non-randomized observational studies to control for measured differences in baseline characteristics. Patients are first grouped in strata determined by their propensity score and then treated and control patients in the same strata are compared directly. Similarly to matching, difficulty arises when the number of baseline characteristics increases.

- **Regression Adjustment based on Propensity Scores:** Propensity scores can also be used in a regression adjustment. Recall, the propensity score is obtained by using a logistic regression model, with exposure to treatment as the dependent variable and all baseline characteristics as independent variables. In regression adjustment, the propensity score is used as the only confounding variable in association with the exposure to treatment (the primary predictor variable) to estimate the effect of the exposure on the outcome.

In the analysis that have been conducted in Chapter 3, the propensity score was used by including it out logistic regression models. However, it was *NOT* the only confounding variable. Variables related to the health condition of the patient have added because was suspected that they were also effecting mortality or effect length of stay.

D.2 Linear Regression

Linear regression is an approach to modelling the relationship between a scalar dependent variable **y** and one or more explanatory variables denoted **X**.

The model of linear regression is:

$$\bar{Y_i} = \beta_0 + \bar{\beta_i} \cdot \bar{X_i} + \mu_i. \tag{D.1}$$

where:
$i \in [1, n]$;
$\bar{Y_i}$ is the dependent variable;
$\bar{X_i}$ is the independent variable;
$\beta_0 + \bar{\beta_i} \cdot \bar{X_i}$ is the regression function;
β_0 is the y-intercept of the regression function;
$\bar{\beta_i}$ is the slope of the regression function;
μ_i is the statistical error.

In linear regression, data are modelled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of **y** given **X**.

D.2.1 Generalized Linear Model

The generalized linear model generalizes linear regression by allowing the linear model to be related to the response variable via a **link function** and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

D.2.2 Logistic Regression

Logistic regression is a special case of generalized linear model with link function as logit function.

$$e(x) = \frac{e(y)}{1 - e(y)} = \alpha + \beta^T f(x), \tag{D.2}$$

It is a regression model applied in cases where the dependent variable **y** is a dichotomous attributable to the values 0 and 1.

D.3 Medical Studies: P-values and Statistical Significance

Null Hypothesis: The independent variable is responsible for random effects (rather than actual difference) in outcome. Whether the difference in outcome is just pure chance based on the effect of this variable. To answer, P-value derivation considers independent variable's coefficient. **The only thing can be said analyzing the P-value is that, when repeating the experiment, in the 97% of the cases a smaller difference between the groups than in the observed ones would be observed, while in the remaining 3% the difference would be greater.**

In this Section the use of P-value in this context will be described. The following contents are drawn from[14]

D.3.1 Null and Alternative Hypothesis

The statistical and probabilistic formalization of medical studies is based on the formulation of a hypothesis to be tested on the basis of collected data. The null hypothesis is that the studied treatment produces absence of effect to the patients or more generally that there is absence of difference between the two treated and untreated patients. The alternative to the null hypothesis (which defines what is expected to be true if the null hypothesis is false), that is that there is a difference between the two groups of patients.

In this context, the statistical and probabilistic formalization of medical studies aim at statistically defining if a given treatment is making or not the difference between the treated and untreated patients in the collected data. Once the data have been collected, their consistency with the null hypothesis will be measured. More precisely, will be determined which of the two hypotheses is statistically more plausible.

The p-value can be used to analyze the importance of a variable in a model. Being in fact the null hypothesis that the inclusion of a given variable x_i in the model does not make a significant contribution, as discussed above a small p-value goes along with the rejection of this hypothesis. Hence, a small p-value, in the already discussed ranges, could indicate if a given variable x_i is important or not for the outcome.

D.3.2 Parametric and Non-Parametrics Hypothesis Tests

A statistical test is parametric if assumes that the data has come from a type of probability distribution and makes inferences about the parameters of the

distribution. In a non-parametric test, instead, the data are not assumed to come from a given distribution.

In the analysis made on the diuretics problem, mortality was studied with the Chi-squared test, a non-parametric test usable for binary (dichotomous) variables, while length of stay was studied with T test, a parametric test usable for continuous variables.

D.3.3 Hypothesis Test with P-value

D.3.3.1 How a P-value is calculated

The P-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. Being $\bar{X}$ the expected value, $\bar{Z}$ the observed value and H_0 the null hypotesis, the P-value is:

$$P(\bar{X} > \bar{Z} \mid H_0) \tag{D.3}$$

D.3.3.2 Interpretation of a P-value

In this context the P-value is defined as the probability that quantifies the strength of evidence expressed by the observed data against the null hypothesis and in favor of the alternative one. In other words, the P-value is a probability that expresses whether it is more plausible that the observed data come from the null hypothesis or the alternative.

A big P-value, more than 0.05, defines that the results on the two compared groups of patients are likely following the same probability distributions and that it is more likely that any obtained difference in the results on these groups is caused by random effects rather then by actual differences. On the contrary, a small P-value, less then 0.05, rejects the null hypothesis, and this means that the differences in the results between the treated and untreated patients are likely to be due to actual differences between the outcomes in the two groups.

Figure D.1 on page lviii shows a graphical visualization of how a P-value can indicate the probability of the null hypothesis.

The p-value is a probability, i.e. between 0 to 1. The levels of significance of the p-value follow:

- **p-value** > 0.1**:** absence of evidence against the null hypothesis.
- **p-value** $\in (0.05; 0.1]$**:** weak evidence against the null hypothesis.
- **p-value** $\in (0.01; 0.05]$**:** moderate evidence against the null hypothesis.

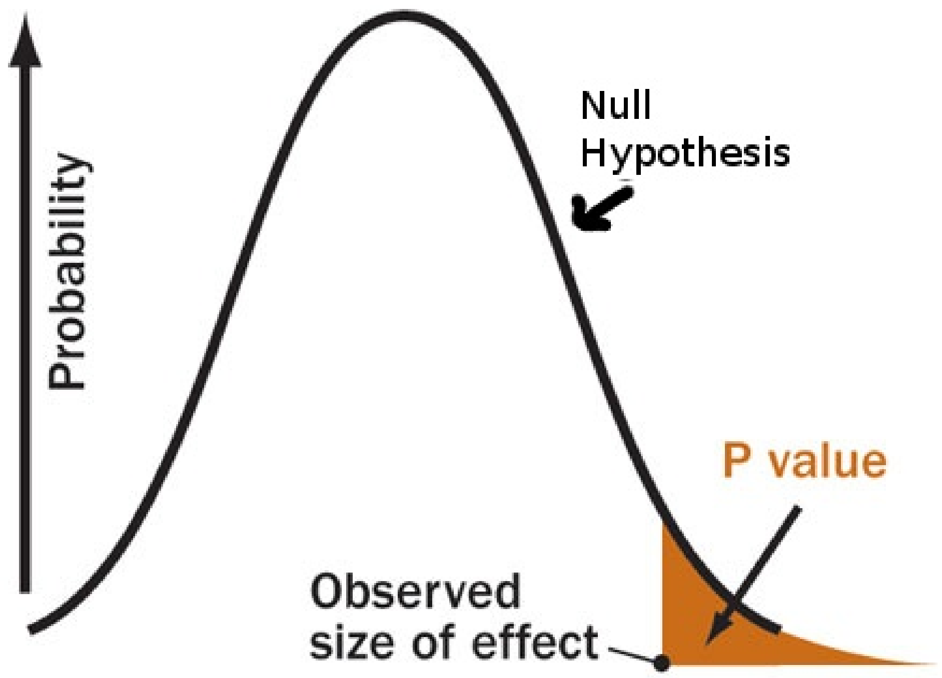


Figure D.1: A p-value can be used to deduce the likelihood of the null hypothesis.

- **p-value** $\in [0.001; 0.01]$**:** strong evidence against the null hypothesis.
- **p-value** < 0.001**:** very strong evidence against the null hypothesis.

D.3.3.3 Clarification in the Interpretation of the P-value

The significance of the p-value is often misunderstood. A p-value equal to 0.03 expresses that there is a 3% probability of observing a difference equal to the one observed in the data even if the means of the two populations are identical, that is even if the null hypothesis is true. One might be tempted to say that there is a probability of the 97% that the observed difference reflects a real difference between the populations and a 3% probability that the difference is due to chance. However, this conclusion would be incorrect: **the only thing you can say analyzing the p-value is that, by repeating the experiment would be observed in the 97% of the cases a smaller difference between the groups than in the observed ones, while in the remaining 3% the difference would be greater.**

Often the p-value is interpreted, wrongly, as the probability that the null hypothesis is true. It is necessary to clarify that a small p-value does not mean that the probability that the null hypothesis is true is lower, but only

that it is more reasonable that the observed data were generated under the alternative hypothesis.
Bottom Line: The only thing you can say analyzing the P-value is that, when repeating the experiment, in the 97% of the cases a smaller difference between the groups than in the observed ones would be observed, while in the remaining 3% the difference would be greater.

Appendix E

Machine Learning

The focus of this Appendix is to describe some methodologies used in the context of problems similar to the one analyzed in this work. Most of the contents are drawn from related papers.

In the first part of the Chapter, will be presented a series of concepts related to knowledge discoverty in clinical databases and further on discribed the methodologies used in similar works.

In the second part of the Chapter, will be provided a background in the context of machine learning and evolutionary computation. In particular the focus will be on describing the Genetic Programming (GP) methodology.

E.1 Knowledge Discovery

In this Section Knowledge discovery will be defined. The contents are drawn from[15].

Knowledge discovery is a concept that describes the process of automatically searching large volumes of data for patterns that can be considered knowledge about the data. It is often described as deriving knowledge from the input data. This complex topic can be categorized according to 1) what kind of data is searched and 2) in what form is the result of the search represented. Knowledge discovery developed out of the Data mining domain, and is closely related to it both in terms of methodology and terminology.

Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Given a set of facts (data) F, a language L, and some measure of certainty C, a pattern is defined as a statement S in L that describes relationships among a subset F_S of F with a certainty c, such that S is simpler (in some sense) than the enumeration of all facts in F_S. A pattern that is interesting (according to a

user-imposed interest measure) and certain enough (again according to the users criteria) is called knowledge. The output of a program that monitors the set of facts in a database and produces patterns in this sense is discovered knowledge[15]. The most well-known branch of data mining is knowledge

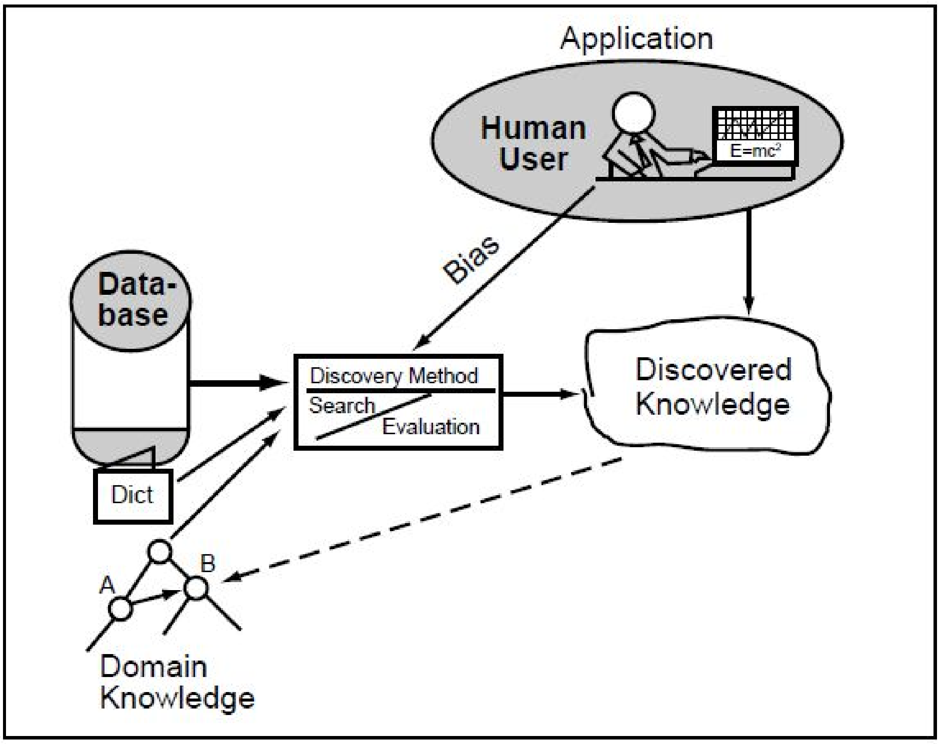


Figure E.1: A Framework for Knowledge Discovery in Databases.

discovery, also known as Knowledge Discovery in Databases (KDD). Just as many other forms of knowledge discovery it creates abstractions of the input data. The knowledge obtained through the process may become additional data that can be used for further usage and discovery.

Although machine learning is the foundation for much of the work in this area, knowledge discovery in databases deals with issues relevant to several other fields, including database management, expert systems, statistical analysis, and scientific discovery.

- **Database Management:** provides procedures for storing, accessing, and modifying the data. Typical operations include retrieval, update, or deletion of all tuples satisfying a specific condition, and maintaining user-specified integrity constraints. The ability to extract tuples satisfying a common condition is like discovery in its ability to pro-

duce interesting and useful statements (for example, Bob and Dave sold fewer widgets this year than last). These techniques, however, cannot by themselves determine what computations are worth trying, nor do they evaluate the quality of the derived patterns. Interesting discoveries uncovered by these data-manipulation tools result from the guidance of the user. However, the new generation of deductive and objectoriented database systems (Kim, Nicolas, and Nishio 1990) will provide improved capabilities for intelligent data analysis and discovery.

- **Expert Systems:** attempt to capture knowledge pertinent to a specific problem. Techniquesexist for helping to extract knowledge from experts. One such method is the induction of rules from expertgenerated examples of problem solutions. This method differs from discovery in databases in that the expert examples are usually of much higher quality than the data in databases, and they usually cover only the important cases, for a comparison between knowledge acquisition from an expert and induction from data). Furthermore, experts are available to confirm the validity and usefulness of the discovered patterns. As with database management tools, the autonomy of discovery is lacking in these methods.

- **Statistics:** slthough they provide a solid theoretical foundation for the problem of data analysis, a purely statistical approach is not enough. First, standard statistical methods are ill suited for the nominal and structured data types found in many databases. Second, statistics are totally data driven, precluding the use of available domain knowledge, an important issue that will be discussed later. Third, the results of statistical analysis can be overwhelming and difficult to interpret. Finally, statistical methods require the guidance of the user to specify where and how to analyze the data. However, some recent statistics-based techniques such as projection pursuit (Huber 1985) and discovery of causal structure from data (Glymour et al. 1987; Geiger, Paz, and Pearl 1990) address some of these problems and are much closer to intelligent data analysis. That methods using domain knowledge is expected to be developed by the statistical community. Statistics should have a vital role in all discovery systems dealing with large amounts of data.

- **Scientific Discovery:** discovery in databases is significantly different from scientific discovery in that the former is less purposeful and

> less controlled. Scientific data come from experiments designed to eliminate the effects of all but a few parameters and to emphasize the variation of one or a few target parameters to be explained. However, typical business databases record a plethora of information about their subjects to meet a number of organizational goals. This richness (or confusion) both captures and hides from view underlying relationships in the data. Moreover, scientists can reformulate and rerun their experiments should they find that the initial design was inadequate. Database managers rarely have the luxury of redesigning their data fields and recollecting the data[15].

E.1.1 Knowledge Discovery in Databases

This and next Sections describe the use of database in the context of knowledge discovery. The contents are drawn from[16] and [15].

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The phrase knowledge discovery in databases was coined at the first KDD workshop in 1989 (Piatetsky-Shapiro 1991) to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the AI and machine-learning fields.

KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data. The distinction between the KDD process and the data-mining step (within the process) is a central point of this article. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns[16].

An example of Knowledge Discovery in Databases Process is shown in Figure E.2 on page lxv.

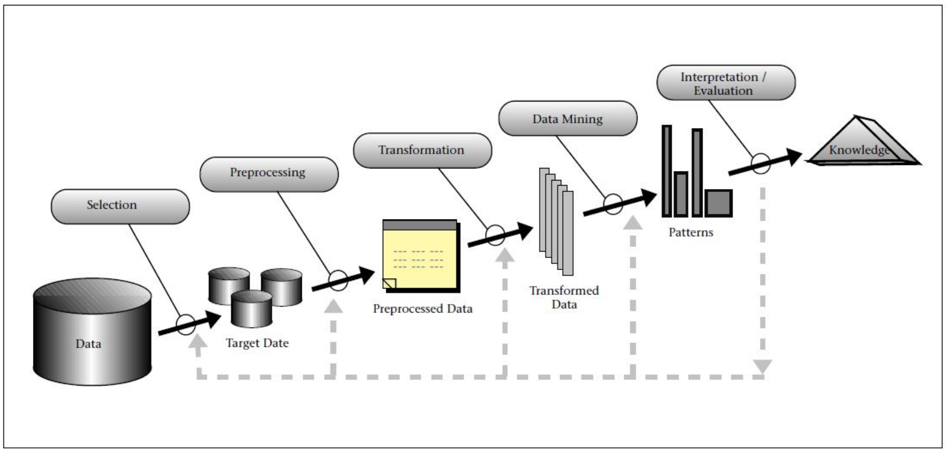


Figure E.2: An Overview of the Steps That Compose the Knowledge Discovery in Databases Process.

E.1.2 Complexity in Knowledge Discoverty

Discovery algorithms for large databases must deal with the issue of computational complexity. Algorithms with computational requirements that grow faster than a small polynomial in the number of records and fields are too inefficient for large databases.

Empirical methods are often overwhelmed by large quantities of data and potential patterns. The incorporation of domain knowledge can improve efficiency by narrowing the focus of the discovery process but at the risk of precluding unexpected but useful discovery. Data sampling is another way of attacking the problem of scale; it trades a degree of certainty for greater efficiency by limiting discovery to a subset of the database (see previous Section on uncertainty)[15].

E.1.3 Clinical Decision Support Systems

This Section discuss the use of computer based techniques in clinical contexts. The contents are drawn from [17] and [18].

Computerized Clinical decision support systems (CDSSs) are interactive decision support systems (DSS)[1] Computer Software, which are designed to

[1] A decision support system (DSS) is a computer-based information system that supports business or organizational decision-making activities. DSSs serve the management, operations, and planning levels of an organization and help to make decisions, which may be rapidly changing and not easily specified in advance.

assist physicians and other health professionals with decision making[2] tasks, as determining diagnosis of patient data.

The goal of diagnosis is to place a nosologic[3] label on a process that manifests itself in a patient over time. However, diagnosis is a complex procedure more involved than producing a nosologic label for a set of patient descriptors. Efficient and ethical diagnostic evaluation requires a broad knowledge of people and of disease states. The nosologic labels used in diagnosis reflect the current level of scientific understanding of pathophysiology and disease, and may change over time without the patient or the patients illness per se changing.

The utility of making specific diagnoses lies in selection of effective therapies, in making accurate prognoses, and in providing detailed explanations. In some situations, it is not necessary to arrive at an exact diagnosis in order to fulfill one or more of these objectives. Treatment is often initiated before an esact diagnosis is made. Furthermore, the utility of making certain diagnoses is debatable. Labeling a patient as having obesity does not flatter the patient, and even worse, may cause the physician to do more harm than good.

In medical diagnostic reasoning, there are also cases where recognition from compiled knowledge does not pertain. Some cases present an overwhelming army of seemingly contradictory information; others present with common conditions in unexpected or unusual manners; some patients manifest rare findings or disorders. Unlike expert chess players who are no better than novices in reproducing random board positions from memory, medical experts have different modes of reasoning that can be invoked when simple pattern recognition based on experience fails. Medical diagnosticians in such settings attempt to reason from first principles, using their detailed knowledge of pathophysiologic processes, to construct scenarios under which an illness similar to the patients might occur[17].

A CDSS allows to match characteristics of individual patients to a computerized knowledge base, and software algorithms generate patientspecific recommendations[18].

There is currently widespread enthusiasm for introducing electronic medical records, computerized physician order entry systems, and CDSSs into hospitals and outpatient settings.

[2]Decision making can be regarded as the mental processes (cognitive process) resulting in the selection of a course of action among several alternative scenarios.

[3]Nosology is a branch of medicine that deals with classification of diseases.

E.2 Machine Learning

Learning is the process of knowledge acquisition in the absence of explicit programming. It can be seen as the process of construction of a program to run a job on the basis of information that do not provide an explicit description of the program itself. Machine learning concerns with the design

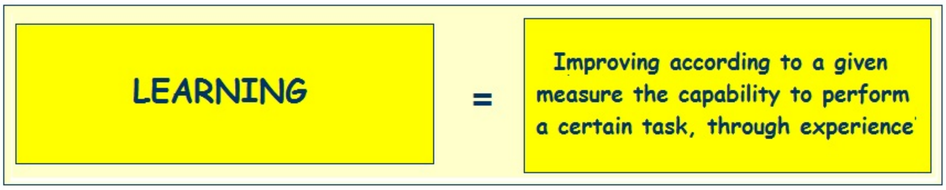


Figure E.3: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E[19].

and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. A learner can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen as examples that illustrate relations between observed variables. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data.

Tom Mitchell in [20] stated that

> Machine Learning is a natural outgrowth of the intersection of Computer Science and Statistics. Could be said that the defining question of Computer Science is how machines that solve problems can be built, and which problems are inherently tractable/intractable? The question that largely defines Statistics is What can be inferred from data plus a set of modeling assumptions, with what reliability?
>
> The defining question for Machine Learning builds on both, but it is a distinct question. Whereas Computer Science has focused primarily on how to manually program computers, Machine Learning focuses on the question of how to get computers to program themselves (from experience plus some initial structure).
>
> Whereas Statistics has focused primarily on what conclusions can be inferred from data, Machine Learning incorporates ad-

ditional questions about what computational architectures and algorithms can be used to most effectively capture, store, index, retrieve and merge these data, how multiple learning subtasks can be orchestrated in a larger system, and questions of computational tractability.

Machine learning, knowledge discovery in databases (KDD) and data mining often employ the same methods and overlap strongly. Infact these fields work with similar basic assumptions: in machine learning, the performance is usually evaluated with respect to the ability to reproduce known knowledge, while in KDD the key task is the discovery of previously unknown knowledge. Evaluated with respect to known knowledge, an uninformed (unsupervised[4]) method will easily be outperformed by supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data.

E.2.1 Complexity of a Problem

The gap between the development of hardware and software technology appears to be one of the biggest unsolved problems in Computer Science. Hardware speed and capabilty has inscreased exponentially during the last few years. Yet an adequate development of software production techniques does not correspond to such a quick and continuous improving in computer hardware performances.

Demand for computer code, more and more efficient and sophisticated, keeps growing in almost every field of industry, but the process of writing code still appears to be slow and obsolete: structured programming, object-oriented programming, and many other techniques allow, today, to write programs in a clean and friendly way, but still each single piece of code is handmade by a *'craftsman'*, the programmer[21].

Hence the attempt to produce techniques that allow computers to learn.

In particular, machine learning methods are already the best methods available for developing particular types of software, in applications where:

- The application is too complex for people to manually design the algorithm. For example, software for sensor-base perception tasks, such as speech recognition and computer vision, fall into this category. All of us can easily label which photographs contain a picture of our mother,

[4]Supervised learning is the machine learning task of inferring a function from supervised (labeled) training data. On the contrary unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data.

but none of us can write down an algorithm to perform this task. Here machine learning is the software development method of choice simply because it is relatively easy to collect labeled training data, and relatively ineffective to try writing down a successful algorithm.

- The application requires that the software customize to its operational environment after it is fielded. One example of this is speech recognition systems that customize to the user who purchases the software. Machine learning here provides the mechanism for adaptation. Software applications that customize to users are growing rapidly - e.g., bookstores that customize to your purchasing preferences, or email readers that customize to your particular definition of spam. This machine learning niche within the software world is growing rapidly.

Viewed this way, machine learning methods play a key role in the world of computer science, within an important and growing niche. While there will remain software applications where machine learning may never be useful (e.g., to write matrix multiplication programs), the niche where it will be used is growing rapidly as applications grow in complexity, as the demand grows for self-customizing software, as computers gain access to more data, and as increasingly effective machine learning algorithms[20] are developed.

E.2.2 Classification Problems

Classification is the problem of identifying which of a set of categories (subpopulations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. The individual observations are analyzed into a set of quantifiable properties, known as various explanatory variables or features. These properties may variously be categorical (e.g. "A", "B", "AB" or "O", for blood type), ordinal (e.g. "large", "medium" or "small"), integer-valued (e.g. the number of occurrences of a particular word in an email) or real-valued (e.g. a measurement of blood pressure). Some algorithms work only in terms of discrete data and require that real-valued or integer-valued data be discretized into groups (e.g. less than 5, between 5 and 10, or greater than 10). An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.).

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term *'classifier'* sometimes also

refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category.

In the terminology of machine learning, classification is considered an instance of supervised learning[5]. The corresponding unsupervised procedure is known as clustering (or cluster analysis), and involves grouping data into categories based on some measure of inherent similarity (e.g. the distance between instances, considered as vectors in a multi-dimensional vector space).

Terminology across fields is quite varied. In statistics, where classification is often done with logistic regression[6] or a similar procedure, the properties of observations are termed explanatory variables (or independent variables, regressors, etc.), and the categories to be predicted are known as outcomes, which are considered to be possible values of the dependent variable. In machine learning, the observations are often known as instances, the explanatory variables are termed features (grouped into a feature vector), and the possible categories to be predicted are classes. There is also some argument over whether classification methods that do not involve a statistical model can be considered *statistical*. Other fields may use different terminology: e.g. in community ecology, the term *'classification'* normally refers to cluster analysis, i.e. a type of unsupervised learning, rather than the supervised learning described in this article.

E.2.3 Evolutionary Computation

The contents of this Section are drawn from [22].

Evolution is any change across successive generations in the heritable characteristics of biological populations. Evolutionary processes give rise to diversity at every level of biological organisation, including species, individual organisms and molecules such as DNA and proteins.

Evolutionary computation uses iterative progress, such as growth or development in a population. This population is then selected in a guided random search using parallel processing to achieve the desired end. Such processes are often inspired by biological mechanisms of evolution.

The principle of evolution is the primary unifying concept of biology,

[5] Learning where a training set of correctly-identified observations is available.

[6] Logistic regression is a type of regression analysis used for predicting the outcome of a binary dependent variable (a variable which can take only two possible outcomes, e.g. "yes" vs. "no" or "success" vs. "failure") based on one or more predictor variables. Logistic regression attempts to model the probability of a "yes/success" outcome using a linear function of the predictors. Specifically, the log-odds of success (the logit of the probability) is fit to the predictors using linear regression.

linking every organism together in a historical chain of events. Every creature in the chain is the product of a series of *accidents* that have been sorted out thoroughly under selective pressure from the environment. Over many generations, random variation and natural selection shape the behaviors of individuals and species to fit the demands of their surroundings.

This fit can be quite extraordinary and compelling, a clear indication that evolution is creative. While evolution has no intrinsic purpose, it is merely the effect of physical laws acting on and within populations and species, it is capable of engineering solutions to the problems of sutvival that are unique to each individual's circumstance and, by any measure, quite ingenious[22].

The most important scientific theory on evolution of species is due to Charles Darwin. He established that all species of life have descended over time from common ancestors, and proposed the scientific theory that this branching pattern of evolution resulted from a process that he called natural selection[7].

Darwin identified a small set of essential elements to rule evolution by natural selection: reproduction of individuals, variation phenomena that effect the likelihood of survival of individuals, heredity of many of the parents' features by sons in reproduction and the presence of a finite amount resources causing competition for survival between individuals.

These simple features , reproduction, likelihood of survival, variation, heredity and competition are the bricks that build the simple model of evolution that inspired the machine learning technique known as evolutionary algorithms (EAs).

An EA uses some mechanisms inspired by biological evolution: reproduction, mutation, recombination, and selection. Candidate solutions to the optimization problem play the role of individuals in a population, and the fitness function[8] determines the environment within which the solutions "live". Evolution of the population then takes place after the repeated application of the above operators.

During the years, many diffent kinds of EAs have been developed. The main feature characterizing the different paradigms of EAs is how the individuals are represented. In particular now will be described the generic algorithms, ancestor of genetic programming.

[7]Natural selection is the gradual, nonrandom process by which biological traits become either more or less common in a population as a function of differential reproduction of their bearers.

[8]A fitness function is a particular type of objective function that is used to summarise, as a single figure of merit, how close a given design solution is to achieving the set aims.

E.2.3.1 Genetic Algorithms

Genetic Algorithms (GA) were invented for the first time by Holland in 1970s, see [23], and later on extended by Goldberg in his works, see [24].

GA's key idea is to adapt the principles of evolution in the way of being able to implement these concepts in a computer so that they can be used to find the solution (or approximate it) for particular problems. In this context, these principles are called genetic operators. Given its nature inspired by natural evolution, many terms used are taken from biology and adapted for this use.

Genetic algorithms, in short, try to simulate the evolution of a species: defined an optimization problem[9], starting from a random set of candidate solutions[10] of the problem, they attempt to improve their quality in an iterative way, applying the genetic operators.

At the beginning of the algorithm is generated a set of possible solutions to the problem, which is indicated with the term of population. Each solution present in the population is codified through a string of bits fixed length and takes the name of individual.

The quality of a solution is instead indicated by the term fitness, which is usually determined by a particular function that takes the name of fitness function.

After the initialization, the evolutionary process starts, which consists of updating at each iteration the set of hypotheses. Each iteration takes the name of generation and it is performed in two phases: the selection process and the variation process.

In the selection phase are calculated the fitness values of all individuals in the population and then is performed a probabilistic extraction on the population based on the values of these fitness. The selected individuals are then used to form the population of the new generation.

In the variation phase are used one or more genetic operators on the individuals of the new population. Not all individuals undergo the process of variation and therefore some of them are replicated unchanged into the next population.

The process ends on the basis of a termination criterion: most com-

[9] It refers to the selection of a best element from some set of available alternatives. In the simplest case, an optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function

[10] A candidate solution is a member of a set of possible solutions to a given problem. A candidate solution does not have to be a likely or reasonable solution to the problem, it is simply in the set that satisfies all constraints.

monly the process ends when at least one individual in the population has a satisfactory fitness or when a prefixed number of generation has occurred.

Follows a deeper discussion on how a genetic algorithm works:

- **Crossover:** is applied to randomly paired strings with a probability denoted p_c. It produces two offsprings, usually different from their two parents and different from each other, but conteining some genetic material from each of their parents. The offsprings are then put into the new population. Many crossover algorithms have been developed for GAs. The most common one is called one point crossover. Its behaviour is shown in Figure E.4. First of all, a number between 1 and

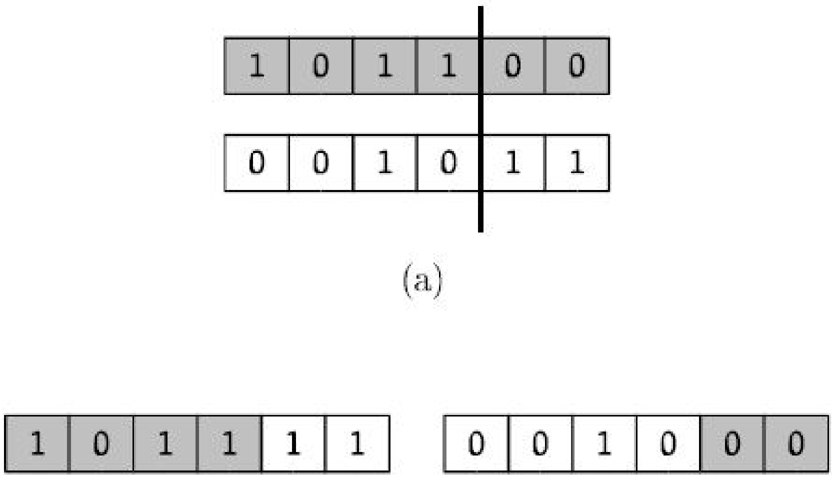


Figure E.4: The GA crossover. Two (a) individuals are selected for crossover. The crossover point, in this case, is 4. In (b) and (c) are shown the results of the crossover process. The two individual are generated by combining the crossover fragment s of the parents.

  $L-1$ is randomly generated using an uniform distribution, being L the length of the individuals' string. This number, that in Figure E.4 is 4, becomes the crossover point. Each parent is then split at this crossover point into a crossover fragment and a remainder. For example, in the picture, the crossover fragment is 1011 and the remainder is 00. The crossover fragment of the first individual is then combined with the reminder of the second one and the crossover fragment of the second individual with the remainder of the first one. The resulting individual are then inserted in the new population.

- **Mutation:** modifies a sub-string of an individual, with a certain probability p_m, and the resulting individual is put into the new population. Many mutation algorithms have been developed. The most commonly used is called point mutation. Its behaviour is shown in Figure E.5. Each position of the current string is chosen with distribution p_m and

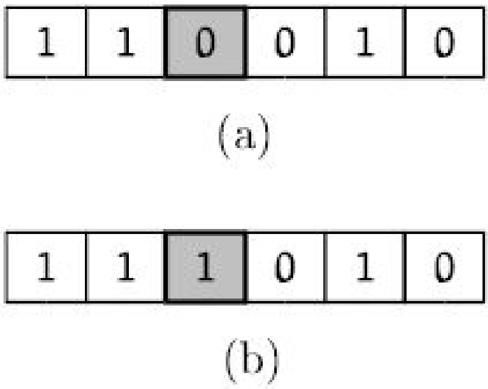


Figure E.5: The GA mutation. In (a) are shown the individuals chosen for mutation with, in this case, mutation point 3. In (b) are shown the individuals resulting by mutation.

  the character contained in that positionis then replaced with another randomly chosen character.

- **The algorithm:** the pseudo-code is shown in Algorithm 4 on page lxxv. The algorithm takes as an input the dimension n of the population, la probability of crossover p_c and the probability of mutation p_m. The output is the best individual till the last population.

E.2.4 Genetic Programming

A lot of the contents of this Section are drawn from [21] and see it for a deeper discussion.

Genetic Programming (GP) is an evolutionary algorithm-based methodology inspired by biological evolution to find computer programs that perform a user-defined task. It is a specialization of genetic algorithms (GA) where each individual is a computer program. It is a machine learning technique used to optimize a population of computer programs according to a fitness landscape[11] determined by a program's ability to perform a given

[11] In evolutionary optimization problems, fitness landscapes are evaluations of a fitness

Algorithm 4 Below is shown and example of pseudocode of a genetic algorithm

Require: $n \geq 0$
  Generation of a random initial population P of n individuals
  while (Termination criterion is not verified) **do**
    Computation of the fitness value for each individual in the population
    Generation of an empy population P'
    while (Population $P' < n$) **do**
      Perform the selection of a pair of individuals x_1 and x_2
      Random extraction of a value r in the interval $[0, 1]$
      if $r < p_c$ **then**
        Perform crossover on x_1 and x_2 optaining their sons y_1 and y_2
      else
        $y_1 \leftarrow x_1$
        $y_2 \leftarrow x_2$
      end if
      Perform mutation on y_1 with probability p_m for each bit
      Perform mutation on y_2 with probability p_m for each bit
      $P' = P' \cup \{y_1\} \cup \{y_2\}$
    end while
    $P \leftarrow P'$
  end while
  Return the best individual of P'

computational task.

The concept of GP was introduced by Koza in [25] and then refined in [26] and [27]. This tecnique aimed at overcoming the fixed length rapresentation of Genetic Algorithms's individuals. This limitetion, infact, is unnatural and contraining for a wide set of applications. For examples, fixed length strings do not readily support the hierarchical organization of tasks into subtasks typical of computer programs, they do not provide any convenient way of incorporation iteration and recursion and so on. But above all, GA rapresentation schemes do not have any dynamic variability: the initial selection of strings length limits in advance the number of internal states of the system and limits what the system can learn[21].

GP, as originally defined by Koza, considers individuals as LISP-like[12] tree structures. These structures are perfectly capable of capturing all the fundamentals properties and features of modern programming languages. The tree-based GP is the oldest and the most commonly used rapresentation, althrough not the only one existing[13][21].

An example of GP individuals is shown in Figure E.6 on page lxxvii.

E.2.4.1 GP Individuals

All the individuals are composed within two groups of symbols: the first one, F, composed by the function symbols $F = \{f_1, ..., f_n\}$ and the second one, T, composed by the terminal symbols $T = \{t_1, ..., t_n\}$. Every function in F has a fixed number of arguments defined arity. Every element of the terminal symbols is a variable or a constant, defined according to the problem. For instance let's consider $F = \{+, -\}$ and $T = \{x, 1\}$, a possible LIPS-like individual could be $(+x(-1 + x))$.

The sets of symbols F and T should have the following properties:

- **Closure:** Every function should be able to take as an argument every possible value of every element of F and T.

- **Sufficiency:** The symbols in F and T should be able to define a solution for the problem. Often the two groups of symbols are not

function for all candidate solutions.

[12]LISP is a family of computer programming languages. It is an expression-oriented language. Unlike most other languages, no distinction is made between "expressions" and "statements"; all code and data are written as expressions. When an expression is evaluated, it produces a value (in Common Lisp, possibly multiple values), which then can be embedded into other expressions. Each value can be any data type.

[13]In particular, in the last few years, a growing attention as been dedicated to linear and graph rapresentations.

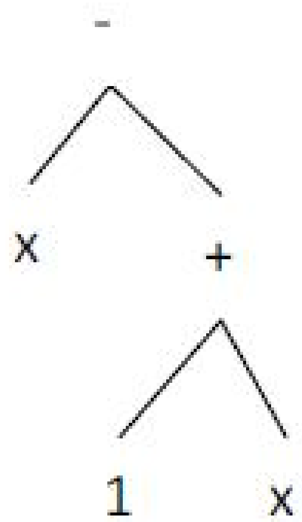


Figure E.6: An example of tree-like GP individual.

known a priori, but decided according to the problem.

E.2.4.2 Initialization of the population

Due to the complexity of the individuals, it is necessary to introduce particular initialization methods of the population. Koza in [27] proposed three methods: *grow*, *full* and *ramped half and half*. For each of those is necessary to specify the set F of functional symbols, the set T of terminals and a maximum allowed depth of the trees.

The *grow* method is performed as follows:

1. Random extraction of a function symbol f_i from F to be used as root of the tree;
2. Being n the arity of f_i, if the current depth is lower then $d-1$, randomly extract n nodes from $F \cup T$ to be used as sons of f_i, otherwise the extraction is performed only on T;
3. Recursively repeat the procedure for all the n extracted nodes.

The *full* method uses the same procedure of the *grow* one, with the difference that extracts the nodes from F when the depth is lower then d instead that extracting them from $F \cup T$.

Both the methods, as observed by Koza in [27], generate a population of trees a lot similar with each others. To avoid this, the method *ramped half and half* has the purpose of preserving divertiry in the population. The idea is to divide the population in d subgroups with the same dimension and assign to each of those a different maximum depth (between 1 and d). Afterwards half of each group is initialized with the *grow* method and half with the *full* one.

E.2.4.3 Fitness Evaluation

Each program in the population is assigned a fitness value, rapresenting its ability to salve the problem. This values is calculated by means of same well defined explicit procedure. The two most commonly used measures in GP are the *raw fitness* and the *standardized fitness*.

The *raw fitness*, as defined by Koza, is "the measurement of fitness that is stated in the natural terminology of the problem itself". It is, therefore, the most natural way to calculate the ability of a program to solve a problem. For instance, if the task is to drive a robot to pick up the maximum number of objects, the *raw fitness* is the number of object picked up by the robot.

The *standardized fitness* restates the *raw fitness* so that a lower value is always better one. Problems where the *raw fitness* is used are also called *maximization problems*, instead problems where the *standardized fitness* is used are also called *minimization problems*.

E.2.4.4 Genetic Operators

For each individual of the in the GP population, three possible actions can be choses: genetic operators can be applied to that individual, it can be copied into the new population as it is, or it can be discarded and replaced by a new individual.

Now will be discussed three genetic operators: *selection*, *crossover* and *mutation*.

The *selection* operator makes the decision of which of the three actions should be applied to the individual. Many possible algorithms have been developed for selection, of those three are the most common: *fitness proportional* (or *roulette wheel*) selection, *ranking* selection, *tournament* selection.

In the *fitness proportional* selection, being N the number of individuals of the population P and $\{f_1, ..., f_{N-1}\}$ their fitness values, each individual has the probability of beeing chosen $p_i = \frac{f_i}{\sum_{i=0}^{N-1} f_i}$. In practice, the probability of being selected is proportional to the value of the fitness.

In the *ranking* selection every individual is sorted according to their fitness values. Every individual is then associated with a function to be chosed according to their rank. This approach was proposed to mitigate the importance of high fitness values in the selection process.

In the *tournament* selection a number of individuals, called *tournament size*, is randomply selected and of those the one with best fitness is chosen.

The *crossover* operator in GP, as in GA, genetates two individuals, y_1 and y_2, from two parents x_1 e x_2. This operator selects a subtree of the individual x_1 and one of x_2 and swaps them, generating in this way two new individuals with genetic material from both the two parent individuals. In Figure E.7 is shown an example of this process. The *mutation*

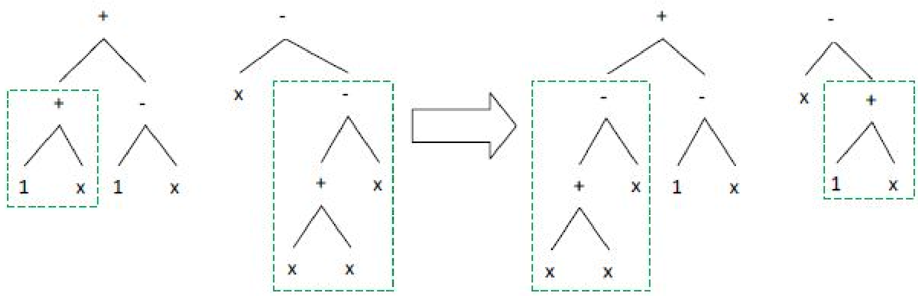


Figure E.7: An example of crossover in Genetic Programming. On the left are shown the parent individuals and on the right the sons. In green are pointed out the swaped subtrees.

operator choses a subtree of an individual and replace it with a randomply generated one. The depth of the new subtree is limited to the maximum depth of the whole tree. An example of mutation is shown in Figure E.8 on page lxxx. Differently from the GA operators, the operators defined in GP are very destructive as they modify the individuals a lot. For this reason, less destructive variants exist. The following techniques aim at this:

- **Steady State:** in this case, after variation, one or two individuals are directly merged into the new population. After the new individuals have been inserted into the population, the new individual are already taken in count for selection. In this way, the steady state works as the variation takes place just one time per generation.

- **Automatically Define Fuction:** some subtrees are considered unchangeble atomic objects. They can be defined ad hoc for the application.

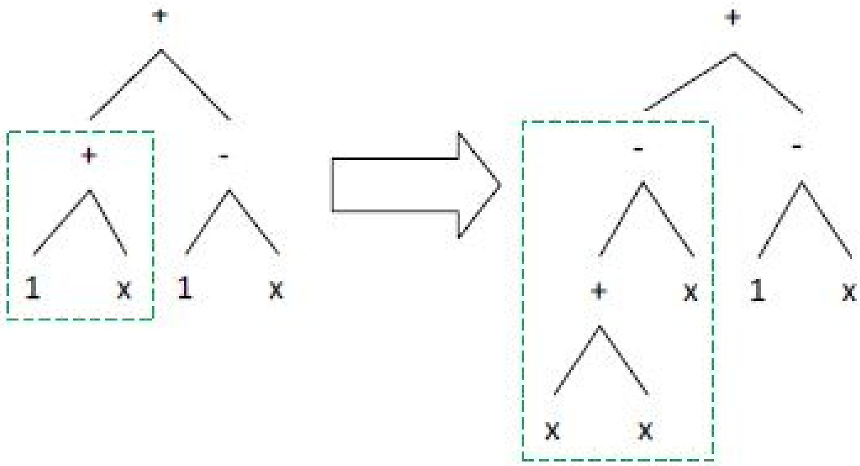


Figure E.8: An example of mutation in Genetic Programming. On the left is shown the original individual while on the right is the shown the tree after the mutation. In green is pointed out the mutated subtree.

E.2.4.5 GP Algorithm

The pseudo-code for the GP algorithm is the same as the one shown in Algorithm 4 on page lxxv for GA.

In synthesis, the GP paradigm breeds computer programs to solve problems by executing the followind steps:

1. Generate an initial population of computer programs (or individuals);
2. Iteratively perform the following steps untill the termination criterion has been satisfied:
   (a) Execute each program in the population and assign it a fitness value according to how well it solves the problem;
   (b) Create a new population by applyng the following operations:
      i. Probabilistically select a set of computer programs to be reproduced, on the basics on thier fitness (*selection*);
      ii. Copy same of the selected individuals, without modifying them, into the new population (*reproduction*);
      iii. Create new computer programs by genetically recombining randomply chosed (*crossover*) parts of two selected individuals;
      iv. Create new computer programs substituting (*mutation*) randomply chosed parts of some selected individuals with new randomly generated ones;
3. The best computer program appeared in each generation is designed as the result of the GP process at that generation. This result may be a soluction (or an approximate solution) to the problem[21].